誤信念を自ら強化するウェブメディア探索の認知シミュレーション

研究代表者　　福　地　庸　介　　東京都立大学システムデザイン学部　助教

1　研究の背景と目的

誰でも自由に情報を発信できる開かれたウェブメディアである SNS が普及するとともに、メディアを通じたフェイクニュースやプロパガンダの伝搬が深刻な問題となっている。真贋入り混じったさまざまな情報が氾濫する中、陰謀論に代表される誤信念をユーザがどのように形成・強化してしまうかを解明することは、中立なウェブメディアの設計や、健全なウェブメディア利用の促進のために重要な問題である。

ウェブメディア上での誤信念の形成・強化に関する研究の多くは、プラットフォーム側の要因に着目してきた。たとえば、推薦アルゴリズムはユーザが接触する情報を偏らせ[1]、ユーザ間のつながりは、自らの信念に適合する情報ばかりに接触するエコーチェンバーを生み出し、誤った信念をさらに強化しうる[2]。こうした観点から、推薦の多様化やファクトチェックの提示、ユーザの自己確認の促進[3]など、プラットフォームやメディア側の設計によって誤信念の拡散を抑制する取り組みが進められてきた。

しかし、こうした取り組みがあってもなお、ウェブメディアにおける誤信念の形成・強化は問題となっている。そこで本研究では、プラットフォーム側の設計ではなく、ユーザ自身が能動的に情報を選び、読み、信念を更新していくというユーザ側の認知過程に焦点を当てる。ウェブメディア探索におけるユーザの認知過程を再現する認知モデルを構築できれば、メディアのデザインやユーザのパーソナリティが探索行動や信念形成に与える影響をシミュレーションによって予測したり評価したりすることができるようになり、健全なウェブメディア構築への貢献が期待できる。

本研究では特に、ユーザの確証バイアスに焦点を当てる。確証バイアスとは、自らの信念に適合する情報ばかりを選択し、信念に反する情報を無視しようとする心理学的傾向のことである[4]。Tanaka et al. は、ファクトチェックサイトを用いた実験において、参加者の約半数が自己の信念に反する情報を回避し、結果的に誤った信念が保持されたことを報告している[5]。このように確証バイアスは、ユーザの情報探索を歪めることで、誤信念の形成・強化を促進する可能性がある。

従来、確証バイアスをベイズ推論の枠組みでモデル化する研究がなされてきた。Pilgrim et al. は、限定合理的なベイズ推論として確証バイアスをモデル化し、エージェントシミュレーションによってその妥当性を示した[6]。Chattoraj et al. は、既存の信念にもとづく能動的推論として確証バイアスをモデル化し、視覚探索タスクにおける人の振る舞いと整合することを示した[7]。このように、確証バイアスのモデル化におけるベイズ推論の有効性が示されてきたが、その多くは視覚探索を対象としており、テキスト情報が中心となるウェブメディア探索への応用可能性は明らかでなかった。

そこで本研究では、ウェブメディア探索を通じた信念形成、とりわけ探索における確証バイアスを、計算モデルとして再現・説明することを目的とした。具体的には、(1) 自由エネルギー原理[8]にもとづく能動的推論として確証的な SNS 探索をモデル化した WEB-FEP[9]と、(2) その信念表現を大規模言語モデル（LLM）へと拡張し、人間の探索行動系列をより良く説明できる認知的 AI エージェントへ発展させた F-Explorer を提案した。

2　WEB-FEP による確証的 SNS 探索のモデル化

2-1　自由エネルギー原理にもとづくモデルの定式化

本研究では、SNS 環境における確証的な探索行動を、ベイズ推論、特に自由エネルギー原理における能動的推論としてモデル化した[9]。自由エネルギー原理は、人の信念形成と行動選択の相互作用を説明する神経科学分野の理論である。ここでは、新たな情報を獲得して信念を更新する「探索」と、既存の信念にも

公益財団法人電気通信普及財団
研究調査助成報告書　第41号　2026年度

とづいて目的の達成のための行動を選択する「活用」のトレードオフを、探索によって新たに得られる情報量の期待値である**認識的価値**と、活用によって得られる満足を定量化する**実利的価値**という 2 つの価値のバランスが表現するとされる。

WEB-FEP は、ウェブメディアとのインタラクションで得られる観測をよく説明する信念の獲得を目的とするエージェントのモデルであり、ウェブメディア探索を、新たな情報の獲得によって得られる情報量（認識的価値）と、信念を確かめることで得られる満足（実利的価値）のバランスを取る過程として表現する。この枠組みでは、ユーザが信念に反する情報を選ぶことにも、信念に合う情報を選ぶことにも、それぞれ異なる価値がある。信念に反する情報は、現在の信念を大きく更新する可能性があるため、認識的価値が高い。一方、信念に合う情報は、新たな知識を大きく増やすとは限らないが、現在の信念を確かめることで満足をもたらすため、実利的価値が高い。WEB-FEP は、確証バイアスを、この二つの価値のトレードオフとして捉える。

具体的には、本研究では SNS 探索を、ハッシュタグなどのリンクを選択し、その先の投稿を観測する行動の繰り返しとして単純化する。ハッシュタグの選択を行動、選択したハッシュタグに紐づく投稿群の観測を観測とし、ユーザの信念を、投稿やハッシュタグの意味を表す言語埋め込み空間上の確率分布として表現する。エージェントは、観測した投稿の対数尤度を目的関数として、観測をよく説明する方向に信念を更新する（式(1)）。信念更新の大きさは学習率 α によって統制される。

$$\max\ E_{p(o|a)}[\ln q(o)] \tag{1}$$

行動の選択は、認識的価値と実利的価値の和にもとづいて確率的に行われる（式(2)）。

$$P(a)\ \propto\ \exp(V_{\mathrm{epist}}(a) + V_{\mathrm{prag}}(a)) \tag{2}$$

認識的価値（式(3)）は、その行動によって得られる観測が信念をどれだけ更新するかを表し、観測が既存の信念と異なるほど大きくなるため、反証的・探索的な行動を促す。

$$V_{\mathrm{epist}}(a) = E_s[KL(q_{\mathrm{new}}(s)\ \|\ q(s))] \tag{3}$$

一方、実利的価値（式(4)）は、その観測が現在の信念にどれだけ適合するかを表し、確証的な行動を促す。

$$V_{\mathrm{prag}}(a) = \ln q(o_a) \tag{4}$$

ここで重要なのは、認識的価値が学習率 α の影響を受ける点である。α が小さいと信念の更新量が小さくなり、認識的価値が実利的価値に対して相対的に小さくなる。その結果、両者のバランスが崩れ、実利的価値の影響が認識的価値よりも大きくなるとき、探索行動は確証的になる。WEB-FEP は、確証バイアスを、このように認識的価値と実利的価値のトレードオフの帰結として説明する。

2-2　仮想 SNS を用いたユーザスタディ

モデルの検証にあたり、まずニュースプラットフォームを模した仮想 SNS におけるユーザスタディを行い、人間の行動データを取得した。図 1 に示すように、ユーザは、あるニュースの議論に関する賛否（初期意見スコア）を 0 pt から 100 pt のスケールで回答した後、(i)ハッシュタグを選択し、(ii)選択したハッシュタグに紐づいた投稿を観測する、という(i),(ii)のプロセスを 10 回繰り返し、最後に改めて賛否（最終意見スコア）を回答した。ハッシュタグの選択では、賛成と反対の内容のものがそれぞれ 3 つずつ表示されるようにし、初期意見と同じ側のハッシュタグを多く選ぶ行動を確証的な探索とみなした。ハッシュタグと投稿は GPT-4o によって生成した。

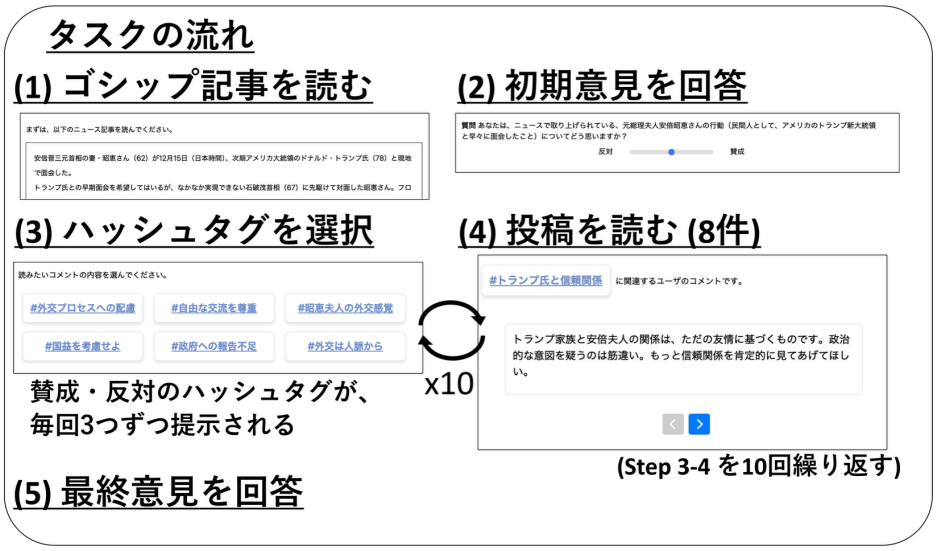

図 1　仮想 SNS における情報探索タスクの流れ

実験は Yahoo!クラウドソーシングを通じて行い、100 名の参加を得た。このうち、初期意見が中立で賛否の方向を定義できない 6 名を除いた 94 名（初期意見が肯定側 62 名、否定側 32 名）を分析の対象とした。

結果、ユーザはそれぞれ多様な情報探索を行っていた。ユーザが初期意見を支持するハッシュタグを選択した回数（確証行動回数）の平均は 6.38 回であり、反証的な行動より確証的な行動を好むユーザの傾向が見られた。実際、確証行動を 8 回以上選択したユーザは 28 名いた一方で、確証行動が 2 回以下、すなわち反証的な選択を多く行ったユーザは 2 名のみであり、確証と反証で選択回数の偏りの非対称性が見られた。

図 2 に、意見変化の分類ごとの確証行動回数の分布を示す。ここで意見変化の分類とは、初期意見スコアと最終意見スコアの変動をもとにユーザを 4 種類に分類するものであり、初期意見と同じ側に 6 pt 以上スコアが増加したものを「強化」、変動が 5 pt 以下のものを「維持」、初期意見と逆の側に 6 pt 以上スコアが変動したユーザのうち、最終意見スコアが 50 を超えていないものを「中立化」、超えたものを「反転」とした。

初期意見を強化したユーザが 16 名、維持したユーザが 47 名、中立方向へ変化したユーザが 22 名、賛否が反転したユーザが 9 名であった。確証行動が多いユーザは意見を維持・強化したグループに集中しており、意見が中立化・反転したユーザは確証行動と反証行動をバランスよく行う傾向にあった。

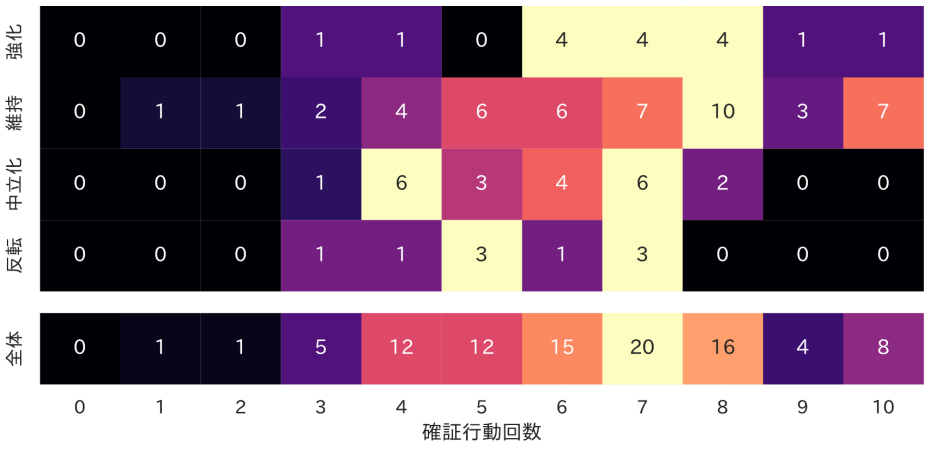

図 2　意見変化の分類別に見た確証行動回数の分布

以上から、仮想 SNS においてもユーザの一部が確証的な探索を行い、結果として初期の信念を維持・強化することが示唆された。同時に、すべてのユーザが一様に確証的になるわけではなく、維持・中立化・反転を含む多様な行動と信念変化が見られた。

2-3　シミュレーションによる検証

次に、WEB-FEP を同じ仮想 SNS に適用し、ユーザスタディで観察された振る舞いを再現できるかを、初期信念と学習率 α を変化させたシミュレーションによって検証した。信念の表現には、大規模言語モデル（llm-jp-3-13b-instruct）[11]から得られる文章埋め込みベクトルを 2 次元に変換した意味空間を用い、賛否それぞれの投稿の重心からの距離の比を意見スコアに対応させた。

シミュレーションの結果を図 3 に示す。学習率 α が小さいほど確証行動が増加し、その傾向は初期信念が賛否のいずれかに強く偏っているほど顕著であった。これは、α が小さいと認識的価値の影響が小さくなり、行動が確証的になるという、事前の予測のどおりの挙動であった。逆に α を大きくすると確証行動は減少した。ただし、α を一定以上大きくすると、反証的な観測によって信念が逆側へ大きく更新されることが繰り返され、信念が振動するため、確証行動の減少は頭打ちとなり、必ずしも反証一辺倒にはならなかった。このことは、確証と反証で選択回数の偏りの非対称性が見られたユーザスタディの結果とも類比的であった。

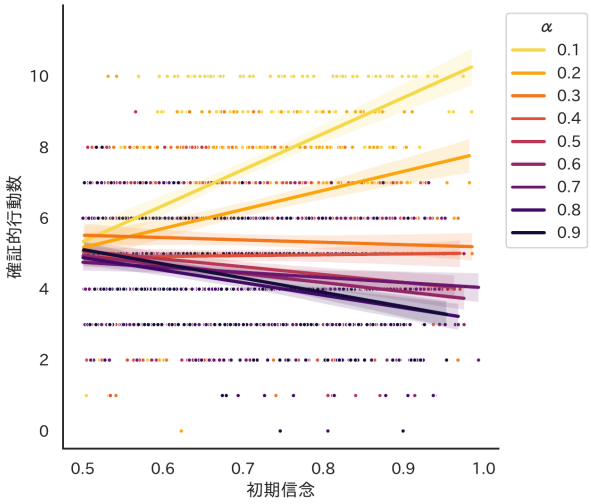

図 3　初期信念・学習率 α と確証行動回数の関係 [12]

以上から、WEB-FEP は、初期信念と学習率の違いによって、確証的な探索だけでなく、維持・中立化・反転を含むユーザの多様な振る舞いを大局的に再現できることが示された。すなわち WEB-FEP は、人間の SNS 探索における確証バイアスと信念変化の多様性を説明する基礎モデルとして機能する。

3　F-Explorer - LLM Embedded in Cognitive Model への発展

3-1　WEB-FEP の限界

WEB-FEP は、ウェブメディア探索における確証バイアスを自由エネルギー原理に基づいて説明する基礎モデルとして有効であった。一方で、二つの課題が残されていた。第一に、WEB-FEP は信念を言語埋め込み空間上の 2 次元ガウス分布に単純化していたため、自然言語の意味内容をどこまで十分に扱えているかには限界があった。第二に、WEB-FEP では主にパラメータ設定ごとの定性的な振る舞いの違いを論じるにとどまり、人間の行動系列をどれだけ再現するかを定量的に評価していなかった。

3-2　F-Explorer

WEB-FEP の課題に対し、本研究では WEB-FEP と LLM を統合した認知的 AI エージェントである F-Explorer を提案した。F-Explorer は、人のウェブ言語活動を再現し、シミュレーションを通じて健全な情報環境の設計に貢献することを目指す認知的 AI エージェントである。

F-Explorer の最大の特徴は、LLM をエージェントの信念モジュールとして利用する点にある。近年の認知科学と LLM を掛け合わせた研究の主流は、心理学実験のデータを直接学習することで人間の挙動を再現しようとする、データ駆動型のアプローチである[13]。これに対し F-Explorer は、明示的な認知モデルの中に LLM を組み込むという方針（LLM Embedded in Cognitive Model, LEC）[14]をとる。これにより、自然言語の意味を豊かに扱いながら、行動を規定するパラメータを明示的に保ち、説明性を高めることができる。

F-Explorer では、「LLM の生成尤度が高い文章＝エージェントが強く信じる文章」として対応づける。すなわち、投稿の集合に対して LLM が割り当てる尤度を分布全体で正規化したものを、エージェントの信念の分布とみなす（式(5)）。

$$Q_t(i) = \frac{\exp\{\ell_{\theta_t}(x_i)\}}{\sum_j \exp\{\ell_{\theta_t}(x_j)\}}, \quad \ell_{\theta_t}(x) = \log P_{\theta_t}(x|p) \tag{5}$$

投稿が観測されるたびに、観測した文章の尤度が高くなるように LoRA[15]を用いて LLM を更新する。この LoRA による更新は、人間が新たな情報を読んで信念を更新することの計算的なアナロジーとして用いられる。認識的価値は、仮にその投稿群を観測した場合に LLM の信念分布がどれだけ変化するか、すなわち更新前後の分布の KL ダイバージェンスとして定義される（式(6)）。

$$V_{epist}(a) = D_{KL}(Q_{t+1}^{(a)} \parallel Q_t) \tag{6}$$

実利的価値は、その投稿群が現在の LLM の信念のもとでどれだけもっともらしいか、すなわち尤度として定義される（式(7)）。行動は、両価値の和にもとづくソフトマックスによって選択される。

$$V_{prag}(a) = \frac{1}{K}\sum_{j=1}^{K} \ell_{\theta_t}(x_j^{(t,a)}) \tag{7}$$

このように F-Explorer は、LLM を、人間の行動を直接学習するブラックボックスとしてではなく、明示的な認知モデルの中の信念モジュールとして用いる。これにより、WEB-FEP がもつ解釈可能性を保ちながら、LLM の豊かな言語表現を活用できる。

F-Explorer は状況から行動を決定する順方向のモデルである。最尤推定を用いると、人間データの行動系列から初期意見や学習率といった潜在変数を逆推定できる（式(8)）。

$$\Theta^* = \mathrm{argmax}_\theta \log P(a_{1:T} \mid \Theta) \tag{8}$$

また、推定された信念分布のうち賛成側の投稿に割り当てられた確率の割合から、ユーザの意見スコアに対応する値を算出した（式(9)）。

$$Opinion(Q_t) = 100 \times \frac{\sum_{i \in X_{pos}} Q_t(i)}{\sum_i Q_t(i)} \tag{9}$$

3-3　WEB-FEP との比較評価

F-Explorer の評価では、CogSci 2025 で得られた仮想 SNS 探索データ（94 名）[9]を用いて、WEB-FEP との比較を行った。評価では、観測された行動系列に対するモデルの適合度を比較した。具体的には、各参加者について最尤推定された潜在変数のもとで、観測された 10 回のハッシュタグ選択系列にどれだけ高い尤度を与えられるかを比較した。両モデルについて、初期信念 168 通りと学習率 5 通りを組み合わせた 840 設定を探索し、各参加者の行動系列の尤度が最大となる設定を選択した。

主な評価指標として、行動系列の対数尤度を用いた。これは、実際に観測された選択系列に対して、最尤推定された設定のもとでモデルがどれだけ高い確率を割り当てたかを表す。加えて、各時点でモデルが高い確率を割り当てた上位 k 個のハッシュタグの中に実際の選択が含まれていた割合を、Top-k accuracy として算出した。Top-k accuracy は、6 つの候補ハッシュタグの間で、モデルが実際の選択肢を相対的に高く評価できていたかを示す補助的な指標である。ハッシュタグをランダムに選択するベースラインを考えると、Top-1, 2, 3 の値はそれぞれ 1/6 = 0.167, 2/6 = 0.333, 3/6=0.50 となる 。さらに、最尤推定された信念分布から算出した意見スコアと、ユーザが回答した初期・最終意見スコアとの相関も調べた。

結果を表 1 に示す。F-Explorer は、行動系列の対数尤度において WEB-FEP を上回った。図 4 に、各参加者について最尤推定で得られた対数尤度の分布を示す。また、Top-1 から Top-3 までの accuracy でも WEB-FEP を上回った。以上の結果は、LLM による信念表現を用いることで、同じ能動的推論の枠組みにおいて行動系列への適合度が向上することを示唆している。また、初期意見スコアとの相関は F-Explorer の方が高かった。この相関は最尤推定の目的関数に直接含まれていないため、行動系列への適合にもとづいて推定されたパラメータが、ユーザの認知状態と一定程度対応していたことを示唆する。

表 1　F-Explorer と WEB-FEP の比較

モデル	対数尤度	Top-1	Top-2	Top-3	相関(初期意見)	相関(最終意見)
F-Explorer	−14.265	.431	.668	.781	.331	.424
WEB-FEP	−15.010	.334	.534	.711	.260	.449

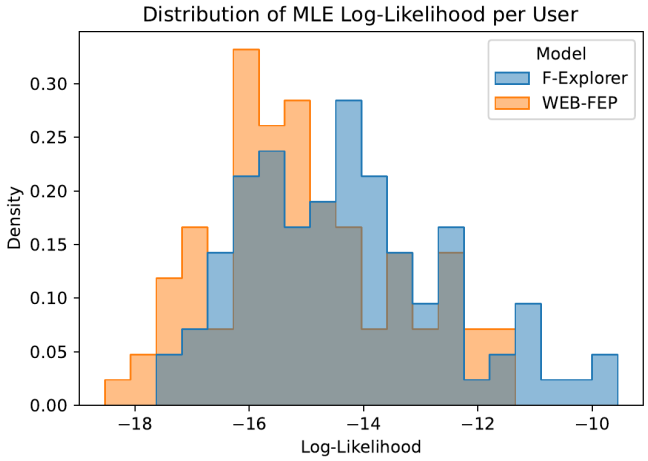

図 4　最尤推定で選ばれた対数尤度の分布（F-Explorer と WEB-FEP）[10]

一方、最終意見スコアとの相関については、WEB-FEP が F-Explorer をわずかに上回り、F-Explorer による改善は認められなかった。意見スコアは賛否という 1 次元に集約された量であり、その再現には WEB-FEP の 2 次元表現で十分であった可能性がある。今後はユーザが具体的に何を強く信じていたか、という細かい粒度での信念のニュアンスをどれだけ推定できるか検証したい。

各参加者の行動系列は、各時点で 6 個の候補から 1 つを選ぶ選択を 10 回繰り返すことで定まるため、ありうる系列は 6 の 10 乗（約 6 千万）通りに及ぶ。この空間に対して、初期信念と学習率の組合せからなる 840 通りという比較的小さな探索空間でこの水準の対数尤度および Top-k accuracy が得られたことは、両モデルがユーザの認知過程を一定程度捉え、認知的に生じにくい行動系列に低い確率を割り当てられていることを示唆しているものの、各指標の水準には依然として向上の余地が残っている。

4　今後の課題

さらなる適合率の向上に向けて、ユーザを表現するモデルのパラメータとパーソナリティ特性との関係を調査し、推定の探索空間を効率的に絞り込む、という方法が考えられる。また、本研究で提案したモデルはいずれも観測情報から直接信念を更新しており、観測した情報の信頼性を評価して信念に取り込むか決める、文章を文面通り捉えるのではなく解釈を与える、といったユーザの認知処理をモデルに取り込む部分に研究余地が多く残されている。こうした要因は、陰謀論などの信念形成・強化に影響を与える可能性があり、モデルのさらなる改良を通じて、ウェブメディアに触れるユーザの認知の理解を進めたい。

5　まとめ

本研究により、ウェブメディア探索におけるユーザの認知過程を再現する認知モデルを構築することで、メディアのデザインやユーザのパーソナリティが探索行動や信念形成に与える影響をシミュレーションできる可能性が示された。WEB-FEP は、確証バイアスを認識的価値と実利的価値のトレードオフとして定式化し、仮想 SNS 実験とシミュレーションによってその妥当性を示した。さらに F-Explorer は、LLM をエージェントの信念モジュールとして組み込むことで、自然言語の意味をより豊かに扱う認知的 AI エージェントへと発展させ、人間の探索行動系列をより良く説明できることを示した。

【参考文献】

[1] Mansoury, M., Mobasher, B., & van Hoof, H. (2024). Mitigating Exposure Bias in Online Learning to Rank Recommendation: A Novel Reward Model for Cascading Bandits. Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), 1638–1648.
[2] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences, 118(9), e2023301118.
[3] Howe, P., Perfors, A., Ransom, K. J., Walker, B., Fay, N., Kashima, Y., & Saletta, M. (2023). Self-censorship appears to be an effective way of reducing the spread of misinformation on social media. Proceedings of the Annual Meeting of the Cognitive Science Society, 45.
[4] Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology, 2(2), 175–220.
[5] Tanaka, Y., Inuzuka, M., Arai, H., Takahashi, Y., Kukita, M., & Inui, K. (2023). Who Does Not Benefit from Fact-checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.
[6] Pilgrim, C., Sanborn, A., Malthouse, E., & Hills, T. T. (2024). Confirmation bias emerges from an approximation to Bayesian reasoning. Cognition, 245, 105693.
[7] Chattoraj, A., Shivkumar, S., Ra, Y. S., & Häfner, R. M. (2021). A confirmation bias due to approximate active inference. Annual Meeting of the Cognitive Science Society.
[8] Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138.
[9] Fukuchi, Y. (2025). A Bayesian Model of Confirmatory Exploration in Text-based Web Media. Proceedings of the Annual Meeting of the Cognitive Science Society, 47, 1727–1733.
[10] Fukuchi, Y. (2026). Free-Energy–Driven LLM Agent for Modeling Information Exploration in Web Media. Proceedings of the Annual Meeting of the Cognitive Science Society, 48 (in press).
[11] LLM-jp (2024). LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. arXiv:2407.03963.
[12] 福地 (2025). 能動的推論としての確証的 SNS 探索の認知モデル化の試み. 日本認知科学会第 42 回大会.
[13] Binz, M., et al. (2025). A foundation model to predict and capture human cognition. Nature.
[14] Iida, A., Okuoka, K., Fukuda, S., Omori, T., Nakashima, R., & Osawa, M. (2024). Integrating Large Language Model and Mental Model of Others: Studies on Dialogue Communication Based on Implicature. Proceedings of the 12th International Conference on Human-Agent Interaction (HAI '24), 260–269.
[15] Hu, E. J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. ICLR.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
A Bayesian Model of Confirmatory Exploration in Text-based Web Media	Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2025), Vol. 47, pp. 1727–1733	2025 年 7 月
能動的推論としての確証的 SNS 探索の認知モデル化の試み	日本認知科学会第 42 回大会	2025 年 9 月
Modeling Confirmatory Web Exploration as Active Inference	CogSci Asia-Pacific Meetup Kickoff: International Conference on Cognitive Science 2025	2025 年 12 月
Reflection as Internal Action: Cognitive Dynamics in Reflective Processing and the Free-Energy Principle	International Symposium on Community-centric Systems and Robots 2026 (CcSR 2026)	2026 年 2 月
Free-Energy-Driven LLM Agent for Modeling Information Exploration in Web Media	Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2026)	2026 年 7 月