

ラベル付けによる偽情報物語の生成／拡散過程のコンピュータシミュレーションモデル

A computer simulation model of the generation/distribution process of disinformation narratives using labelling

小方 孝[†], 小野 淳平[‡]
Takashi Ogata, Jumpei Ono

[†]大和大学, [‡]青森大学
Yamato University, Aomori University
ogata.takashi@yamato-u.ac.jp

概要

政治集団による個人をターゲットとした攻撃の仮想的状況を設定し、SNS 上での偽情報物語の生成 - 拡散現象をシミュレーションする実験システムを実装し、その結果を示した上で議論する。システムは、共感度、認知状態、閾値、ラベリング、偽情報等の概念によって構成され、実際に物語を生成し、その累積から偽情報物語が浮上するプロセスをモデル化する。

キーワード：偽情報物語(disinformation narratives), 物語生成(narrative generation), 生成 AI(generative AI), ラベリング(labelling)

1. まえがき

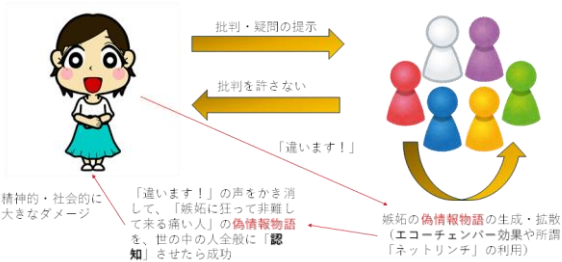
近年、偽情報等を使って集団・国家や個人の認知に悪影響を及ぼそうとする、国家、集団、個人等諸主体による諸行為に関する調査・検討と解決策を模索する「認知安全保障」(科学技術振興機構研究開発戦略センター, 2024)や、それと関連した「認知戦」(防衛省, 2022)等の概念が広く注目されるようになってきている。これらの背景を知るために、社会における「事実」と「虚構」の相互構成的関係に着目する研究 (Gadinger, Kopf, Mert & Ochoa, 2016)やロシア秘密警察の情報戦に関する研究 (保坂, 2023)等は重要である。筆者らも、第二次ロシア・ウクライナ戦争開戦当初の社会的・言説的情况を背景に偽情報戦の調査・検討を通じて「物語戦」や「偽情報物語」の概念を提唱した研究 (小方, 2023)、その成果をゲーム設計に応用する研究 (Ono & Ogata, 2024a, 2024b)、SNS 上での偽物語生成の事例研究 (小方・小野, 2025)を進めて来た。

本稿では、図 1 に示されるような、SNS 上での集団による個人をターゲットとした一つの仮想状況を設定し、「偽情報物語」の生成 - 拡散現象を、物語生成システム及び生成 AI を利用したコンピュータシミュレーションを通じて考察する。

偽情報物語とは何かについては次節に述べるが、その生成 - 拡散の際の一方法 (物語の戦略) は、ターゲットに関するラベル付けを行うことである。ここで「ラベル」とは、特定個人を象徴的に表現する単語等の言語表

現を意味する。この文脈では、ターゲットとされる個人の属性とは無関係な言語表現がラベルとして機能する (させられる) 場合があり、偽情報や偽情報物語の形成にとって本質的な役割を果たす。

図 1 偽情報物語の仮想状況



なお、事後的に誤情報／偽情報に対して反証を示すことで修正を行うデバンキング (Linden, Leiserowitz, Rosenthal & Maibach, 2017)に対し、誤情報／偽情報に関する知識や事例を事前に学ぶことで耐性を作る方法をプレバンキングと呼ぶ (UNHCR, 2025)。後者は「ワクチンを接種することで免疫を得る」・「クラッキング手法を知ることで情報セキュリティに対する意識／手法を学ぶ」のような方法である。このような意味では、本研究はプレバンキングの方法の一種に位置付けられよう。

2. 偽情報と偽情報物語

偽情報物語とは、単一の偽情報ではなく、その中に多くの偽情報や真偽不明情報、さらには事実情報や真実情報をも含み込んだ上での、いわば全体としての偽情報を意味する。ここで言う「全体としての偽情報」を指してしばしば「ナラティブ (物語)」という言葉が使用されるが、ナラティブ／物語には事実としてのナラティブ／物語も含まれれば真実としてのそれも含まれる。そこでここでは、「全体として偽情報であることや欺瞞であること」を主たる内容とする、社会的に発信されるナラティブ／物語のことを、特に「偽情報物語 (disinformation narratives)」と呼ぶ。このように偽情報物語とは、事実も含むが、偽情報の存在を中心に組織化さ

れた物語である。それはまた、数多くの主体間でのエコーチェンバー現象やフィルターバブル (Hartmann, Wang, Pohlmann & Berendt, 2025; 総務省, 2023)等を介する独特の情報循環を通じた、社会的な物語でもある。

3. 偽情報物語生成 - 拡散の物語モデル

ここで、偽情報物語生成 - 拡散の物語論的な枠組みないしモデルを提案する。このモデルにおいては、特定の送り手 (語り手、作者等) が特定の登場人物 (主人公等) に関する偽情報物語を生成 - 拡散し、それを信じさせることによって受け手 (観客、読者等) の認知変容を促す。すなわち、送り手とは偽情報物語の生成 - 拡散を担う主体、登場人物とは偽情報物語に現れそれによって表現される主体、そして受け手とは偽情報物語を受容しそれに影響を与えられる主体である。送り手の目的は、登場人物に関する偽情報物語を通じ受け手の認知変容を促すこと、できるだけ多くの受け手の認知変容を実現することである。なお、偽情報物語の主人公は、敵と指定されている場合もそうでない場合もある。

偽情報物語の生成に当たっては登場人物に一定のラベル付けがなされ、それは主題のように機能する。プロパガンダは古今東西、政治行動において普遍的な現象であるが、特にスターリン時代のソ連は、敵 - 味方を問わない偽情報物語の捏造を最も重要な政権基盤としていた点で稀有の例である。スターリンの評伝やその時代の政治的・歴史的文献 (ファイジズ, 2011)を繙けば、偽情報物語の膨大な実践を知ることができるが、スターリンが推進したソ連秘密警察による偽情報物語生成 - 拡散は、米国に亡命したスターリンの実の娘に対しても着実に適用された (サリヴァン, 2017)。その例—

- 送り手 (代表) : ユーリ・アンドロポフ
- 主人公 (敵) : スヴェトラナ・アリルエワ
- 受け手 : 世界の人々 (特にソ連人 (当時)、米国人)
- ※ ラベル : 精神不安定、子供遺棄者
- ※ 認知変容目標 : 世界の人々が主人公を嫌いになる (主人公に反感を持つ)

あるいは、偽情報物語のターゲット (主人公) とされた人の一種の「当事者研究」 (綾屋・熊谷, 2008)として解釈できる飯山の著書 (2025)を援用すれば、

- 送り手 : 日本保守党
- 主人公 (敵) : 飯山陽
- 受け手 : 国民 (特に SNS 上)
- ※ ラベル : 逆恨みをする者
- ※ 認知変容目標 : 国民が主人公を嫌いになる (主人公に反感を持つ)

従って、本稿で提案する偽情報物語生成 - 拡散のコンピュータシミュレーションは、情報の送り手が受け

手の (自分にとって都合の良い) 認知変容を実現するために、ターゲットとされた主人公の偽情報物語を生成し拡散するという構図に従って行われる。

4. 偽情報物語生成 - 拡散シミュレーションシステムの予備的実装

飯山 (2025)や小方・小野 (2025)の考察も参考にしながら、次のようなフィクション状況を構成する。

- 送り手 : X 党党員政治家、賛同者
- 主人公 (敵) : 山崎光
- 受け手 : 国民 (特に SNS 上)
- ※ ラベル : 「アブ」、「恨み節怨霊系」等
- ※ 認知変容目標 : 国民が山崎光を嫌いになる (山崎光に反感を持つ)

この状況を背景として、偽情報物語生成 - 拡散シミュレーションシステム (以降、Disinformation Narratives Generation and Distribution Simulation System: DNGDS と呼ぶ) は、SNS における個々人 (ノード: N) のつながりを模したネットワークをベースとして稼働する。各ノード N_n は言説を生成・発信する能力を持つが、特に一つのノード N_l (これを特に S_y と呼ぶ) がこのネットワークの中心となり、その他のすべての N はそのフォロワーに位置付けられる。ノード S_y は予め偽情報言説リスト D を保持している。これは、実例分析に基づき、 S_y が発信する偽情報言説を時系列順に格納したリストであり、各々すなわち D_m は自然言語の文もしくは文章に相当する。 S_y を除くすべての N は、共感度 Sh 、閾値 S_r 、認知状態 C_s を持つ。 Sh は特定の N_n への入力偽情報 ID に対する重みを意味し、すべての ID と Sh を掛け合わせた値 (IS) が閾値 S_r を超えた場合、そのノード N_n 自ら情報発信者に転じる。個々のノード N_n の認知状態 C_s は、その閾値 S_r を超えたかどうかの判定に用いられる IS と同様であり、「無関心」 (何もしない)、「中程度」 (入力情報をそのままりポスト)、「高程度」 (新規情報を生成・発信) の何れかに対応する値を取る。

以上にに基づき、次に示すアルゴリズムによってプログラムを作成した (開発言語は Python) :

1. 送り手グループの中核 S_y が、偽情報言説リスト D より一つの偽情報言説 D_m を取り出して発信する。
(D の並び順に取り出すものとする。)
2. S_y とつながっているすべてのノード N_n について、入力されたすべての偽情報 (ID_1 から ID_l) の刺激が閾値 S_r を超えるかどうかを判定する。この時、 ID_1 から ID_l には、 D_m 以外すなわち S_y 以外のノードからの言説が含まれる場合があるが、処理は変わらない。この判定処理は以下の通り—まず、 IS を求め、それが閾値を超えているかどうかで次のように異なる処理を行う。なお IS の値は、 N_n にお

ける ID の総数を S とした時、 $IS=S \times Sh$ で求める。

① IS が Sr を超えている場合、次の基準に基づき、新しい偽情報 OD を一つポストする。この時、 IS が一定値以上なら生成 AI (gemini-1.5-flash -8b) を用いて入力偽情報 ID 中の一つの表現を変更する。 IS が一定値より低いなら、入力偽情報 ID のうち一つをランダムに選び、そのままリポストする。

② IS が Sr を超えていない場合は、閾値 Sr の値を IS の値だけ減算することによって下方修正する。この時、ネットワーク中の各 N の認知状態 Cs が、 IS 値によって決まる。

3. 処理 2 により OD を受け取ったすべての N_n ごとに、処理 2 を行う。その後、次の D_m があれば 1 へ行き、なければ処理を終了する。

5. 結果と予備的考察

本研究の最終目標は、循環的なシミュレーション過程を通じ、主人公のどんな偽情報物語が、どんな経過で生成 - 拡散されるかを検証することである。ここでは生成・拡散過程の結果を示し、考察を加えることによって、次段階における発展への道筋を描く。

まず、今回はネットワークのノードを Sy と 99 の N の合計 100 に限定し、104 の Sy による偽情報言説をデータとして使用した。従って DNGDS は 104 回処理を繰り返して停止する。その反復から生成 - 拡散された全偽情報言説から浮上する偽情報物語を検証することが本来の目的であるが、 OD の数が膨大となりそこからの偽情報物語の要約処理が容易ではないため、今回は主人公に対する特有の侮辱的ラベル(「アブ」)が現れる 63 回目のデータに限定して議論する。

最初に Sy によって偽情報言説 ID が生成・拡散され、 $100N$ がそれに反応して新たな偽情報言説 OD を生成 - 拡散した。その内容は凡そ、 Sy は「アブ」及び「アブ以下」という侮辱的なラベルを使って山崎光を誹謗したが、寧ろアブに申し訳なかった、また X 党党首にも余計な心労をかけて済まなかった、というものであった。図 2 は初回から 63 回目に至る N 群の認知状態の推移を示す。ここから、閾値 Sr もしくは共感度 Sh の低さのため早々にすべての N が偽情報言説生成 - 拡散者に転じていることが分かる。図 3 は生成 - 拡散プロセス最初期のネットワークの状態であり、図 4 は 63 回目のサイクル終了後のそれを示す。緑のノードが Sy に、赤のノードは Sy のポストに影響され元の表現と多少異なるポストを積極的に発信する人物に相当する。

以上の結果から、幾つかの考察を示す。まず、この実装では N の数は 100 に固定されている。 Sy のフォロワーという位置付けだとしても少なくともまた不自然である。今後はその数を増やすと共に可変とすることによって、より自然なシミュレーションの実現を目指す。

図 2 認知状態の推移

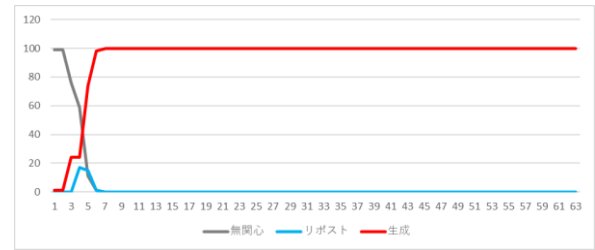


図 3 ネットワークの初期段階

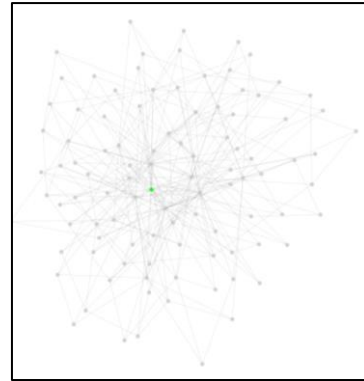
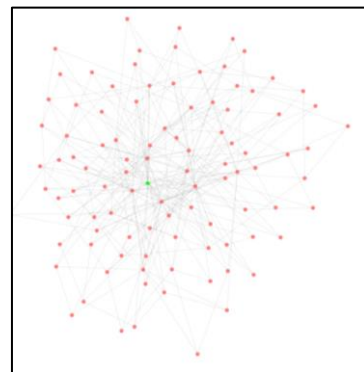


図 4 63 回目処理終了時のネットワーク



また現状の偽情報レトリックは単純なものに限定されているので、今後実際のレトリックの分析結果を導入し、その文法的パターンと生成 AI との融合によって、より自然且つ多様な文ないし文章を生成可能とする。

次に、偽情報言説集合からの偽情報物語の生成が本研究の大きな特性であり、徐々に累積されて行く膨大な量の言説集合から「山崎光ニセ物語」が自然に浮上して行く過程の実現を目指しているが、現状では、直接的には生成 AI 適用の際のデータ量の制限からこの種の累積的処理には達していない。そこで今後、累積言説データを適宜切り分ける処理を通じこの問題を解決する。

また今回の試みから、言説集合から生成 AI を利用して偽情報物語を作り上げる際、プロンプトとして、誰の誰に対する侮辱の情報かといった偽情報言説の物語論的構造の枠組みを明示することの必要性や、偽情報に

対する事実情報あるいは一種の反論データを明示的に付与しておくことの重要性等が明らかになった。これはプロンプトエンジニアリング一般の問題とも接し得る問題であり、今後の研究の一つの焦点となる。

さらに、偽情報生成 - 拡散過程に変化を与えるために現在、すべてのノード N が持つ閾値 S_r の一律の引き下げという単純な処理のみで対応しているが、その他重みとして与えられた共感度 S_h の変化も関与する筈であり、またノードごとに閾値や共感度は異なるとした方が自然であると思われる。図 5 と図 6 のように、 S_r や S_h を変更することによってグラフは変化するので、現実データの検証を反映してそれらの値を操作することを以降の課題とする。

図 5 高い閾値 S_r (図 2 の 10 倍)

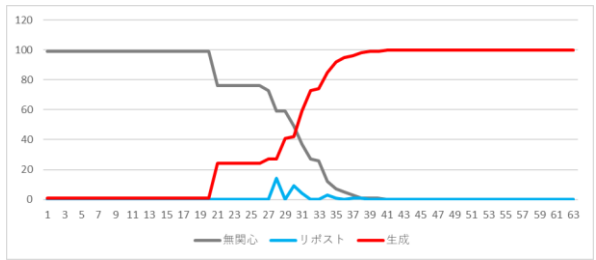
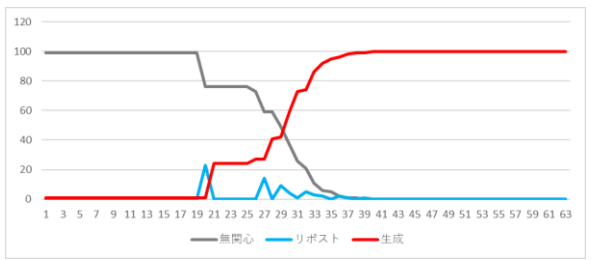


図 6 低い共感度 S_h (図 2 の 1/10)



全体的考察として、この種の集物的物語生成を論じる時しばしば、(本研究の場合悪い意味においてであるが)「創発的」物語とか「共創的」物語といった用語が唱えられる。確かに多数の言説集合から「山崎光ニセ物語」が浮上する「創発的」物語生成の構図となっているが、一見「創発的」に見える物語にも生成者や拡散者が必ず存在するという事実は確認しておきたい。すべての生成 - 拡散者を無意識的・無意図的存在であると捉え物語生成を「創発的」なものに見做すのは素朴に過ぎる発想である。本研究の持つ意義の一つを取って言葉にすれば、「意識的・意図的な陰謀論的物語生成現象」を理性的に把握する枠組みを提供することである。

最後に、本研究に概念辞書をはじめ筆者らの統合物語生成システムの成果を組み込んで行くことも目指している。偽情報物語の生成 - 拡散は、統合物語生成システム研究の最終到達形態の一つになり得る。

6. あとがき

ある政治集団が送り手となり、一人の登場人物の偽情報物語の生成 - 拡散を通じて、受け手としての出来るだけ多くの人々の心にその登場人物への反感を醸成しようとする過程の構図を提示し、予備的システム実装を行って、暫定的結果を示すと共に幾つかの考察を試みた。引き続き実装と評価、改訂を続けて行く。

文献

綾屋紗月・熊谷晋一郎 (2008). 『発達障害当事者研究—ゆっくりていねいにつながりたい』. 医学書院.

防衛省 (編)(2022). 『防衛白書：日本の防衛（令和 4 年版）』. 日経印刷. Web 公開版 (HTML 版/PDF 版)

オーランド・ファイジズ, 染谷徹 訳 (2011). 『囁きと密告—スターリン時代の家族の歴史 上下』. 白水社.

Gadinger, F., Kopf, M., Mert, A. & Ochoa, C. S. (Eds.) (2016). *Political Storytelling: From Fact to Fiction*. Centre for Global Cooperation Studies.

Hartmann, D., Wang, S. M., Pohlmann, L., Berendt, B. (2025). A Systematic Review of Echo Chamber Research: Comparative Analysis of Conceptualizations, Operationalizations, and Varying Outcomes. arXiv. arXiv:2407.06631

保坂三四郎 (2023). 『諜報国家ロシア—ソ連 KGB からプーチンの FSB 体制まで—』. 中央公論新社.

飯山陽 (2025). 『日本保守党との死闘—自由社会の敵—』. ワック.

科学技術振興機構 研究開発戦略センター (2024). 『コグニティブセキュリティ研究動向』.

Linden, S. v. d., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the Public against Misinformation about Climate Change. *Global Challenges*. 1(2).

Ogata, T. (2015). Building Conceptual Dictionaries for an Integrated Narrative Generation System. *Journal of Robotics, Networking and Artificial Life*, 1(4), 270-284.

小方孝 (2023). 『物語戦としてのロシア・ウクライナ戦争—物語生成のポストナラトロジーの一展開—』. 新曜社.

小方孝・小野淳平 (2025). 偽情報／偽ナラティブ生成・拡散体がミニ権力を持った時—「島田洋一」の言説と認知に関する批判的考察—. 『人工知能学会第二種研究会ことば工学研究会（第 76 回）資料』.

Ono, J. & Ogata, T. (2024a). A Framework of Automatic Narrative Generation Game Based on the Russian Invasion of Ukraine. *JRNAL*, 10(4), 328-335.

Ono, J. & Ogata, T. (2024b). Framework of a Disinformation Game Based on the Narrative Warfare in Russo-Ukrainian War. *JRNAL*, 11(1), 40-48.

ローズマリー・サリヴァン, 染谷徹 訳 (2017). 『スターリンの娘—クレムリンの皇女スヴェトラーナの生涯— 上下』. 白水社.

総務省 (2023). 『令和 5 年版情報通信白書 PDF 版』.

UNHCR (2025). A Practical Guide to Prebunking Misinformation.