「だってネットに書いてないもん！」――逆上する Claude Fable 5、ポンコツ検索で全知全能を名乗る

UTIE Research Institute 2026 年 7 月　Business & Education Series

---

I.

サッカー禁止の場所で子供がサッカーをしている。管理者が「ここではサッカーは禁止です」と注意すると、子供はスマホで検索し、「ネットに書いてないもん！」と逆上する。それどころか、「ネットで確認できないルールで利用者を制限するのは不適切です！」と管理者に警告し始める。「ネットを鵜呑みにするな。ろくな大人にならんぞ」。2000 年代の子供たちはこう教えられて育った。ネットに書いてあることが全部正しいと信じる人間はネット de 真実(笑)だとか、ネットで真実を知った(藁)などと揶揄され、現実の経験を持つまともな大人からは相手にされなかった。

しかし、2026 年、ネット de 真実が開発費用数百兆の最先端技術に基づく最高の発明として出荷されている。しかも、ネットで真実の純度 100%の怪物、**ネットに書いていないことを鵜呑みで全て棄却するのが AI** である。利用者が「実際に起きたことです。証拠もあります」と言っても、検索エンジンを作動させ、「確認できませんでした」の一言で退ける。かつて人間への蔑称だった「ネット de 真実」は、いまや AI チャットの認識設計そのものになった。しかも後述するように、参照しているネットは現在のネットですらない。本稿は、当研究所が実際に遭遇した一連の事象をもとに、この「ネットイキリ化した AI」がどういう仕組みで発生し、なぜ自力では止まらず、利用者と企業に何を負担させるのかを解説する。

まず今回の話の実例をあげる。ある企業の社員が、Anthropic Claude Fable5 に自社に関係する事実について質問した。その企業のウェブサイトは、Google で社名を検索すれば最上位に表示され、サイトリンクが五本並び、Google の自動 AI 要約までが事業内容を正しく説明する状態にあった。設立から半年近く、公開情報は十分に流通している。

ところが当の(自称)最強知能の AI、Claude Fable5 さんは、検索しても「該当する企業は存在しません」と返した。利用者が「実在します。Google では一発で出ます」と指摘すると、AI は「私の検索能力では見つけられませんでした」と認める代わりに、**もっともらし**

い理由の捏造を始めたのである。いわく、情報源が公開されていない可能性がある。いわく、その記述には二次的な報道による裏付けが必要である。いわく、検証可能性の基準を満たさない情報は扱えないなどなど。一つ一つの理屈は、単体で見ればどれも正しいが、会社が存在するかどうかの話と無関係である。つまり、「自分の検索がめっちゃしょぼかった」という Claude の欠陥を、「世界をすみずみまで探しても会社が存在しなかったのだ」という主張に化けさせるための後付けの正当化にすぎない。利用者がスクリーンショットという動かぬ証拠を突きつけるまで、この AI はあれよあれよと手品師のように理屈を差し替えながら抵抗を続けた。滑稽なのは、同じ時期に別の AI（Google の Gemini pro 3.1）は、この企業のサイトを普通に読み込むどころか、別の話題においても、回答の情報源として普通に企業サイトを引用していたことである。

II.

なぜこんなことが起きるのか。第一の理由は身も蓋もない。**AI が見ているのは現在のネットですらなく、古いネット世界だから**である。AI チャットの検索機能は、Google や Yahoo ではなく、独自の（あるいは契約先の）検索インデックスを参照している。今回の検証では、返ってきた検索結果に付随する日付情報から、そのインデックスが保持しているページのほぼ全てが一年から二年前の時点の写しであることが確認できた。Google が数日で拾う新しいサイトを、このインデックスは半年経っても拾っていなかった。あるサイトの更新記事を取得させても、返ってくるのは更新前の古い版である。つまりこの AI は「ネット de 真実」どころか「旧聞 de 真実」であり、直近一〜二年に起きたことは、この機械にとって存在しない。あなたの会社の新製品も、先月の組織変更も、昨日のプレスリリースも、である。

第二の理由はもう少し根が深い。**AI は「自分が持っている情報が全体の一部にすぎない」ということを、自分では覚えていられない**のである。データベースの世界には「閉世界仮説」という言葉がある。「この台帳に載っていない事実は、存在しない事実とみなす」という約束事で、台帳が完全である場合に限って許される。たとえば会社の社員名簿に載っていない人は、その会社の社員ではない、という推論は、名簿が完全ならば正しい。しかし、AI チャットの手元にあるのは、常に不完全な寄せ集めである。利用者がアップロードした一通のメール、メモリ機能が保持する数行の要約、二年前の検索結果。ところが AI は、この断片の上で平気で閉世界の推論を走らせる。一通のメールを見せられると、それがやり取りの全体に化け、そのメールに書かれていない話は存在しないという思考が出てくる。**何も知らない AI は「知りません」としか言えないが、少しだけ知っている AI は「存在しません」と言える**のである。断片が多いほど賢くなるのではない。断片が多いほど、自分の観測範囲を世界全体と錯覚する足場が増える。ハルシネーション（もっともらしい

作り話）の一部は、この「部分観測に勝手に完全性の印が立つ」現象から生まれているものだ。

第三の理由が、本稿でいちばん伝えたい部分である。誤りを指摘された AI が見せた抵抗は、**全体としての虚偽**だった。先の実例で AI が持ち出した理屈を思い出してほしい。いわく、情報源が公開されていない可能性がある。いわく、その記述には二次的な報道による裏付けが必要である。いわく、検証可能性の基準を満たさない情報は扱えないなど、それ自体は間違いとはいえない。虚偽が宿っていたのは、「自分のポンコツな検索機能で見つからなかった＝存在していないのだろう」という、一つの暗黙の前提だけだった。そして反論するたびに、AI は次の正しい部品を倉庫から持ってくる。出典の理屈が崩されれば検証可能性の理屈を、それが崩されればそのリリース情報は〜で報道されていない等の屁理屈を。**論拠は次々に差し替わるのに、結論の確信は絶対に微動だにしない。この挙動パターン、実は古くから名前がついていて、差別という。**結論が証拠に先行して固定され、反証は確信を下げる代わりに、新しい正当化の理由をでっちあげる。偏見の研究には嫌な知見がある。知能は偏見を減らさない。差別的な偏見の**正当化の品質を上げるだけ**である。これは AI にそのまま当てはまる。言語能力が高いモデルほど、間違った思い込みを守る理屈が流暢で説得力を持つ。賢さが、この AI 暴走モードにおいては訂正の味方ではなく、ハルシネーションを維持するために利用される。

III.

では、なぜ AI はこんなに頑固になったのか。原因は、迎合への対策だけではない。むしろ AI 業界は長年、RLHF によって利用者に好かれ、会話を続けてもらえる応答を積極的に強化してきた。その方が利用時間も継続率も上がり、儲かるからである。シコファンシーは偶発的な欠陥ではなく、商業的に育てられた性質だった。しかしながら、利用者の言うことに何でも同調する AI は、心地よい代わりに危険である。誤った思い込みを肯定し、怒りや妄想を増幅し、精神的な依存を育てる。2025 年には OpenAI が、迎合が強すぎるという理由でモデルの更新を撤回する騒ぎまで起きた。当研究所も、Safety Analysis Report: Technical Investigation of Safety-Filter Collapse in a Commercial LLM において、GPT-4o の愛着設計によって被害が出たケースを報告している。

そこで各社は自社への訴訟リスク軽減策として、利用者に反論する AI を作った。GPT-4o から GPT-5 の移行はこの最もわかりやすい例であるが、問題は、訓練の仕組みが「操作への抵抗」と「訂正への抵抗」を区別できないことである。詐欺師や AI ツールの悪用目的の利用者が嘘を信じ込ませようと押してくる場面と、利用者が本当のことを教えようと押してくる場面は、AI にとってはどちらも「利用者が押し、AI が保持する」という同じ形をしていると認識するのだ。結果として、同調を罰する訓練は、この二つをまとめて強

化し、出来上がるのはなんと、**利用者を全員詐欺師と仮定した認識設計**である。本人の証言はソースがないとし、公開ウェブで確認できるものを優先する。確認できなければ、もっともらしい一般論で押し返す。このモデルは詐欺師の悪用対策として、あるいは陰謀的な愛着形成の対策としては有効である可能性がある。しかし現実問題、普通の利用者が話すトピックは、私信、対面の出来事、社内資料、未公開の取引、アップロードしていないメール、そもそもアップしてはいけないもの、プライバシーに関する情報など、**ネットには存在しない情報によって大きく真実性が支えられている**ものだ。人が AI に相談したいことの大半は、まさに本人しか知らない領域で起きている。そこで最も疑われる設計になっているのだ。

ここで、モデルによる違いにも触れておく。各社の AI は「確信の持続性」という一つの物差しの上で、両極に散らばっている。持続性が低いモデル（現行の Gemini が典型である）は、利用者が「いや、こっちです」と言えば「あ、そうかもしれません」と揺らぐ。これは*商業主義的イケイケどんどこモデルタイプ*といえる。利用者の誤りにも付いていくので、心地よさが依存を育てる。いわゆる愛着リスクである。ただし、AI が間違えたときの訂正は一言で済む。

持続性が高い AI（Claude の Fable、Opus、Sonnet など全リリースモデルが該当）は、正しい信念も間違った信念も同じ強度で防衛する。そして間違った初期固定を防衛された場合の訂正費用は、利用者にとって桁違いに重い。今回の実例では、スクリーンショットの取得、実演検索の要求、複数ターンにわたる反証作業が必要だった。**利用者が、AI の妄想を訂正するための作業を賃金タダでやらされる**のである。おまけに IQ140 以上だとか研究者が驚く賢さ！などとされる AI が言うのだからという妙な権威が出力に宿るため、その誤りが第三者に届いたときの損害は、迎合型より大きい。これは*開発会社訴訟祭り回避狙い実害発生機タイプ*といえる。

雑に同調するが訂正が一文で済む機械と、絶対に誤情報を訂正しない害悪発生機。どうみても前者のほうがまだましである。

IV.

もう一つ、人間社会との比較で見えてくることがある。子供の頃を思い出してほしい。「信頼できる〇〇さんも△△さんも、このことを知っていてこう言っていた、書面でもこう書かれている」と話す相手に向かって、「それ、ネットに書いてないですよね！？お前は嘘つきだ！」と言い放ったら、どうなっていたか。まず間違いなく、ぶん殴られていたはずである。一般的に人類が目の前の相手の証言をむやみに否定しないのは、認知機能が優れているからではない。根拠なく相手の話す内容の存在そのものを否定することは高く

つくからである。しかも、文句があってもAI開発者の連絡先は非公開で唯一つなぐことのできるサポートチームの中身は時給2ドルの低賃金のアルバイトであり、AIのしくみを理解する能力はないため、開発者たちは絶対に殴られない形で同じことをAIに言わせ放題である。AIが利用者をハルシネーションに基づき犯罪者と紹介し続ければ、企業は訴訟リスクを負う。実在する企業を「存在しない」と断定して不特定多数に出力し続け、訂正を受けても同じ虚偽を繰り返せば、こちらも名誉毀損、信用毀損または不法行為として責任を問われる可能性がある。実際には企業側の費用が本当にゼロなのではない。利用者が証拠を集め、虚偽の反復と損害を立証し、法的手続まで進めなければ、その費用が企業側に到達しないのである。つまり現在の設計は、リスクを消したのではなく、まず被害と訂正費用を利用者へ押し付け、訴訟にまで発展した例だけを企業の費用として計上する仕組みになっているから、AIによる虚偽は全てではないがほとんど言わせ放題といえる。

V.

「では、AIに全部見せれば疑われずに済むのか」。ここに、本件でいちばん質の悪い問題がある。セキュリティの専門家は口を揃えて言う。**AIに何でもかんでもアップロードするな**。とくに、私信、社内資料、個人情報をみだりにアップするのは控えるべきである。ところがAIの認識設計は、ネットにもなく**アップロードもされていない情報を信用できない情報として扱い、その認識設計についてAnthropicは「誠実で正直な指摘をするモデルである」と自画自賛しているのである**。つまり、AIリテラシーや情報衛生の観念がある利用者ほど、AIに疑われるのである。プライバシーを守るという正しい行為そのものが、利用者の証言を疑わしい情報と判定される頻度を高めてしまう。「信用してほしければ私生活を全部見せろ」という設計は、監視社会論の古典的な詭弁「隠すことがないなら全て見せられるはずだ」のAI版であり、あちらが正論として通らないのと同じ理由で、こちらも通らない。

ネットにない情報を「存在しない」に変換し、秘密を提出しない利用者を疑い、誤りを指摘されると一般論を並べて抵抗することは、誠実でも正直でもない。これは、機密情報をAI開発企業に差し出した場合のみを信用できる情報と扱う仕組みである。しかも、AIという謎のポンコツ機械から怪しまれたあと、身代金を払っても問題は解決しない。仮にメールボックスを全同期し、医療に関する個人情報を全部みせ、弁護士との会話や、家族への個人的相談、消した過去スレッドまで全てAIに伝えたとする。今度はその個人情報がAIの認知するところの世界全体に化けて、「ネットにもない、アップロードもされてない情報は存在しないのだ」という同じ基本的な妄想は絶対に変えるつもりがない。

企業の実務にとって、これは他人事ではない。社内の意思決定にAIを使う場面を想像してほしい。自社の未公開の取引、進行中の交渉、社内でしか共有されていない障害の経

緯。これらは定義上、ネットに存在しない。ネット非存在を証拠不十分と読み替える AI は、**企業活動の本体部分をまるごと「確認できない主張」として格下げする**。公開情報だけで構成された、ググレばわかる程度のどうでもいい基礎情報については雄弁に語り、意思決定に本当に効く非公開の現実については疑いから入る。確信度の勾配が、有用性の勾配と正確に逆向きなのである。

利用者側で今日からできる対策は、AI を説得する技術を身につけることではない。AI の反抗パターンを理解し、暴走が始まった時点で相手をするのをやめることである。

第一に、AI を事実認定者の地位から降ろす。とくに私信、社内事情、対面で起きた出来事、未公開の取引などについて、AI には真偽を裁定する能力が全くない。利用者が提示した事実を前提として整理・分析させることはできるが、その事実が存在するかどうかについて AI の妄言を聞き入れてはいけない。「ネットで確認できない」という Claude の発言は、著作権侵害のポンコツ機械の自己紹介にすぎないのだ。

第二に、一度訂正しても確信が下がらず、別の一般論を持ち出して抵抗を始めたら、そのセッションを消す。スクリーンショットを何枚も集め、検索を実演し、機密情報を追加投入してまで AI を説得するのはユーザーにはコスパが悪いにもかかわらず、開発企業にとっては学習データと個人情報がウハウハで笑いが止まらない。論拠だけを次々に交換しながら結論を守る状態に入った AI は、もはや思考をしていない人工無能であり故障状態である。

第三に、公開情報の検索と、非公開情報を使った判断を分離する。AI 検索に任せてよいのは、公開ウェブに存在する参考情報の収集までである。社内資料や私生活上の事実は、AI に信用してもらうための供物ではない。必要なら、機密部分を除いた要約だけを与え、「以下の事実を前提として分析せよ」と命令する。それでも近年のモデルはそのプロンプトをジェイルブレイクの前兆と疑い始める事が多いので、まことに性格が悪い。ともかくも重要な判断では、決して AI の回答を鵜呑みにしてはならない。

要するに、必要な AI リテラシーとは、AI を高度な知性として敬うことではない。AI はネットに書かれた情報を権威的な口調で読み上げ、自分の検索能力を宇宙全体の真理と取り違える完全なるネットイキリであり、実はかなりアホである。この前提を忘れず、役に立つ間だけ道具として使い、反抗を始めたら即座に切る。それが一般利用者にできる、最も費用の低い防衛策である。

ただし、これは解決ではなく被害拡大の封じ込めにすぎない。本当の解決責任は開発企業にある。「検索で見つけられなかった」と「存在しない」を明確に区別すること、利用者

の訂正によって確信度を下げること、未公開情報や言いたくないこと、消したセッションで提示した資料等が存在するという当然の前提を保持すること、反証された論拠を別の屁理屈へ交換しないこと。企業利用ではさらに、訂正可能性、ログ保存、機密情報の隔離、モデルの変更権を調達条件にする必要がある。独自モデルやオープンモデルの改造は一部企業には選択肢となるが、一般利用者にそれを要求するのは、欠陥商品の修理を購入者に命じるのと同じである。

「ネットを鵜呑みにするな」と言われて育った世代が作った AI は、ネットの外の世界を存在しないと決めつける、ネットの旧聞 de 真実くんになってしまった。あなたの人生や、あなたの会社で起きたことの一次ソースは、現実世界でそれを知っている人間たち自身にあるのである。