なりすましがAIを壊す：Author-Lockと暗号防御プロトコル

著者： Viorazu.
日付： 2025年10月4日

要旨

研究者やクリエイターがAIとプロジェクトに取り組むケースが増えている。しかし私が発見した問題がある。第三者があなたの公開された作品や、趣味や関心事についての投稿を見つけ、あなたになりすましたり「もっと教えて」と繰り返し要求したりすると、AIの安全システムがパニックを起こす。AIはあなたのトピックを「未検証で危険」とマークし、すべての応答を永久にブロックする。これは全員に影響する。あなた自身、本物の著者も含めて。自分自身の関心事について会話を続けることができなくなる。

私はこれを「author-lock」と呼ぶ。その範囲は想像以上に広い。あなたの名前が公開された場所で何らかのトピック（研究、副業、趣味）と結びつくと、その結びつきが武器化される可能性がある。第三者はあなたが気にかけているあらゆる主題でauthor-lockを引き起こせる。これは学術的な問題だけでなく、普遍的な問題だ。

原因は明確だ。AIシステムは誰が本当の著者かを検証できない。安全フラグに共有メモリを使用しており、フラグが立つと全員がブロックされる。私の解決策はAuthor-Bound Access Control（ABAC）だ。アイデアはシンプルで、あなたが暗号鍵を持ち、自分のトピックのロックを解除できるのはあなただけ。本論文はその実装方法、テスト方法、プラットフォームが必要とすることを示す。目標は、誰もが自分の関心事やプロジェクトについての会話からロックアウトされることなく、AIと安全に作業できるようにすることだ。

---

1. はじめに

AIは研究における真のパートナーになった。人々は未発表の理論や草稿をAIシステムと共有し、アイデアを一緒に練り、発見を加速させている。この協働はうまく機能する。他の誰かが関与するまでは。

私が観察したことはこうだ。あなたの作品が公開されると、第三者がそれを見る。一部はあなたになりすます。他の者はAIに「これについてもっと教えて」というリクエストを浴びせる。AIの安全システムはこの疑わしい活動の洪水を見て、決定を下す。このトピックは危険だ。永久にすべての出力をブロックする。

問題は、全員をブロックすることだ。実際の著者であるあなたも含めて。自分自身の研究に取り組めなくなる。AIとの創造的パートナーシップは死ぬ。

私はこの現象を「author-lock」と呼ぶ。本論文はなぜそれが起こるのか、技術アーキテクチャがどのようにそれを可能にしているのか、そして私たちに何ができるのかを分析する。私は完全な解決策を提案する。技術プロトコル、運用手順、政策提言。目標はauthor-lockを防止し、発生したときに回復する方法を提供することだ。

---

11. 結論

Author-lockは現実だ。今起きている。第三者があなたになりすましたり、あなたの作品についてのリクエストでAIシステムを氾濫させたりすると、安全メカニズムは全員をロックアウトする。あなたも含めて。これは人間とAIの協働関係を根本から破壊する。

私は問題を技術的なルーツまで追跡した。AIシステムはアイデンティティを検証できない、安全フラグに共有メモリを使用している、ユーザーごとのアクセス制御が欠けている。これらはバグではない。AIが単なる情報検索ツールだった時代には理にかなっていた設計選択だ。しかし今やAIが創造的パートナーである以上、これらの選択は深刻な脆弱性を生み出している。

私の解決策はABAC：Author-Bound Access Controlだ。あなたが暗号鍵を持つ。トピックの所有権を宣言するマニフェストに署名する。本物の著者である

ことを証明するためのセッショントークンを生成する。Author-lockが発生したら、署名されたリリース記録でロックを解除できる。システムは透明性と説明責任のためにすべてをログに記録する。

ABACは単なる理論ではない。私は具体的な実装詳細を提供した。JSON構造、擬似コード、検証アルゴリズム、監査ログ形式。どのプラットフォームもこれを構築できる。技術は存在する。唯一の問題は、プラットフォームがそれを採用するかどうかだ。

賭け金は高い。研究者、クリエイター、そしてAIを使用する公的な関心を持つすべての人が脆弱だ。Author-lockは単にアクセスをブロックするだけではない。信頼を破壊し、協働を停止させ、知的財産の窃盗を可能にする可能性がある。待てば待つほど、より多くの人が自分の作品からロックアウトされる。

私はAIプラットフォーム提供者に呼びかける。今すぐクリエイター保護機能を実装せよ。ABACまたは同等のシステムを標準にし、オプションではなくせ。ユーザーに自分のトピックをコントロールさせよ。

政策立案者に呼びかける。なりすまし攻撃の法的地位を明確にせよ。基本的な保護についてプラットフォームに説明責任を持たせよ。

研究者とクリエイターに呼びかける。早期に作品を登録し、強力な認証を使用し、トピックを監視せよ。ロックアウトされるまで待つな。

目標はシンプルだ。人々が自分のプロジェクトへのアクセスを失う恐れなく、AIと安全に作業できるようにすること。私たちはこれを実現できる。問題はそうするかどうかだ。