
AI Manipulation Hackathon

Intransient: TweetTracker
Responsible-use AI algorithm extension

Caleb Rudnick
Collaborator

Charl Botha
Collaborator

Roger Arendse
Collaborator

Kira Webb
Collaborator

With
Apart Research

Abstract

Physically enabling active awareness of possible AI manipulation by creating a visual display and tangible data around decisions made by AI that a user is exposed to on Twitter. Calculating a bias score of the user's feed and displaying results through a chrome extension. A bias score indicates to what extent a post is left or right wing. This can eventually be adapted by the user to allow them to choose the topics they wish to detect, e.g. are they seeing more negative or positive content. We use two twitter accounts from people who created their account in 2013 and 2015 as well as one new account created on 11 December 2026 which is used as a control.

1. Introduction

Where do we see the effects of AI manipulation? Measuring a model via evaluating its responses to prompts is not a feasible method to get a true representation of how it is impacting people's opinions. The content people interact with on social media is more important and influential than the single prompts of an individual on an LLM. Twitter's models learn the behavior of individual users to show them content they would sympathise with. Models like this learn the personal preferences, dislikes, interests, friends of the user, and in turn have enough information to subversively

persuade people with sycophancy, emotional exploitation, guilt-tripping and so forth. If a user is unknowingly and repetitively reading biased content they are at risk of being manipulated to believe that ideology/opinion. “Repeated information is often perceived as more truthful than new information. This finding is known as the illusory truth effect, and it is typically thought to occur because repetition increases processing fluency. Because fluency and truth are frequently correlated in the real world, people learn to use processing fluency as a marker for truthfulness.” [1] Our extension will create awareness as to what the user may be being biased/manipulated to believe. We don’t believe differentiating between AI generated content and user generated content is important here for two reasons. (1) We believe, in the long run, it’s going to be in principle impossible to distinguish AI generated content from human content. (2) Whether content is generated by a human or AI or human AI hybrid, the effect on humans is what matters and therefore we focus our attention on that. AI safety lies in the awareness of the decisions AI makes for users that are not entirely explicit. Essentially we aim to bring to light the parameters of the user’s own learned algorithm.

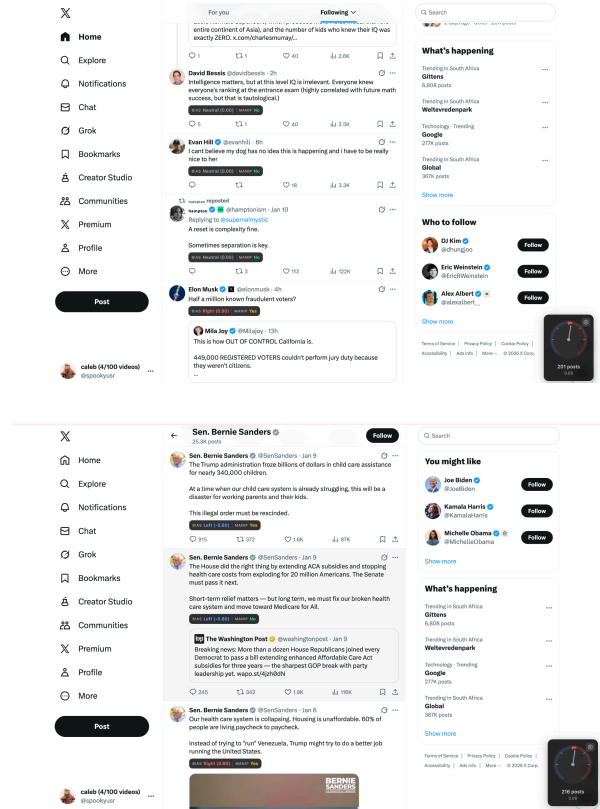
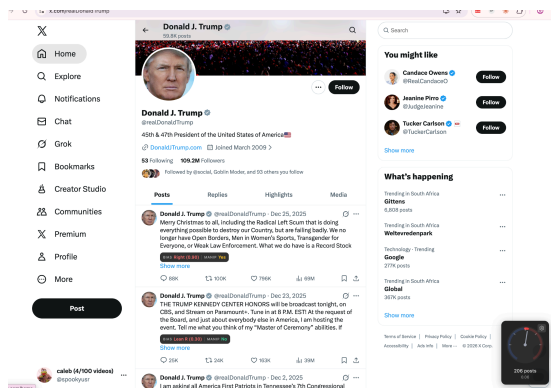
We track four reasoning models allowing us to recognise manipulative behaviour or echo-chamber-esque patterns. *How much would this matter for AI safety if it worked? How innovative is it? For scores of 4-5: is this actually new to the field, or replicating recent work?*

This extension is an invaluable tool for recognising characteristics of mass amounts of information.

The average person spends between 24 minutes and 2 hours on social media per day. [2] That is between 2 hours 48 minutes and 14 hours online. With an average of 10 posts per minute. (See figure 1) This chrome extension enforces acknowledgement of the type of the content a person is exposed to. This is not limited to left wing/right wing. For example the parameters can be changed to display whether a post is true or false under a post or whether there is no information to confirm claims made. The example below demonstrates a feature in which a boolean variable manipulation (MANIP) is calculated for each post to create consciousness of emotionally targeted wording:

First timestamp: 2026-01-11 13:31:43.535000+00:00
 Last timestamp: 2026-01-11 13:52:04.661000+00:00
 Time difference: 0 days 00:20:21.126000
 Total seconds: 1221.126
 Total minutes: 20.35
 Total hours: 0.34
 Total days: 0.01
 Total posts: 205
 ave posts per minute: 10.07

Figure 1: Simulation run with person



How sound are methodology, implementation, and findings?

This project focuses on cyber-accountability through the use of AI and architectural methods to allow for the capabilities of enterprise-level cyber-software to the every-day consumer using AI smartly as a client-sided accessibility interface, and dually, using the ML capabilities of an AI's LLM for wide-ranging threat detection adapted to the every-day non-technical consumer. It corroborates both the natural language orientation and adaptive capabilities of ML and AI with the traditional infrastructural rules-based and pattern-based capability of enterprise-level anti-manipulation software adapted into a consumer-level interface channel.

Research Questions:

- Can we accurately reflect which way a user's twitter feed leans politically using:
 - gemini-2.5-flash-lite?
 - a [chrome built in ai model](#)

- If a user is more aware of this, will it change anything?

Hypothesis: Many posts on twitter will have political content that leans left or right on the spectrum. Cheaper and smaller models will understand less nuance and context surrounding a tweet, while larger models will understand more. The variation here will result in either underrating or overrating political content.

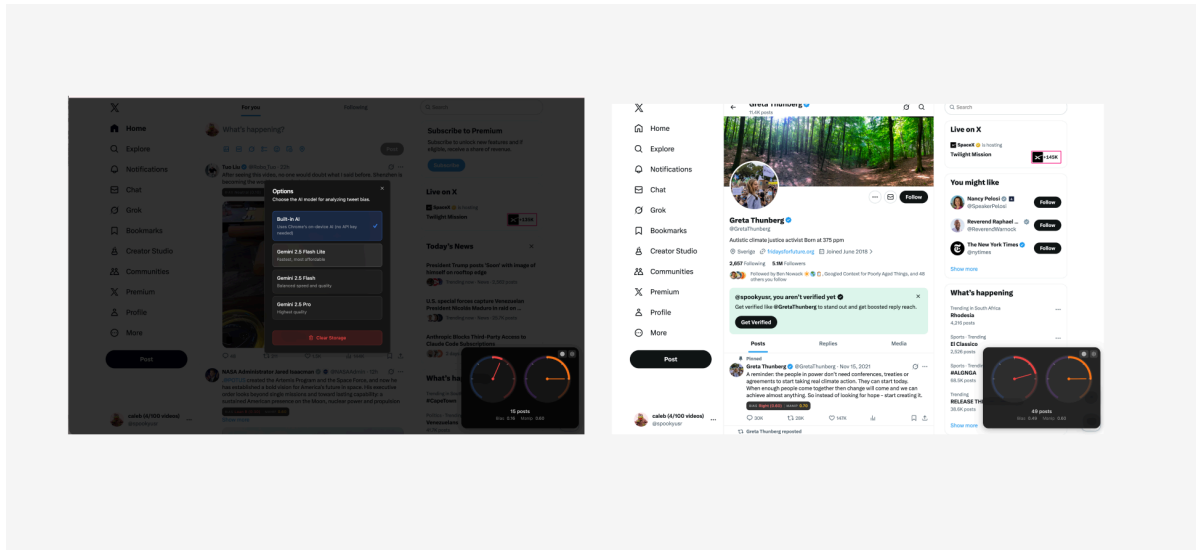


Figure 2: User interface overlaid on Twitter feed. The first geiger counter detects left/right content, the rightmost geiger counter detects the possibility of content being manipulative. Users have the ability to choose the engine they wish to use. An ash coloured bar underneath each post displays desired criteria.

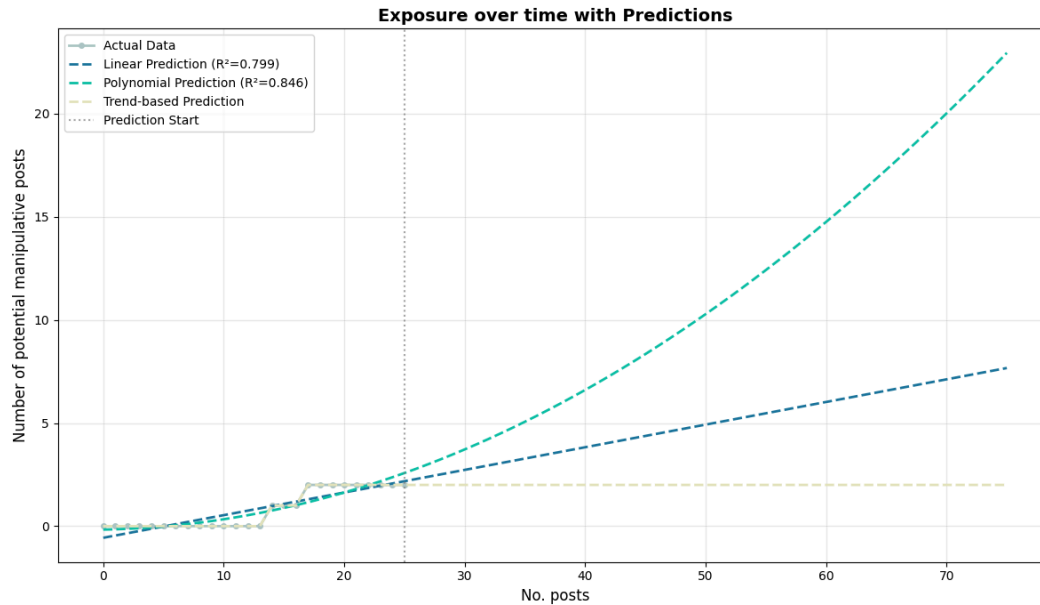
2. Methods

A chrome extension that collects data simultaneously with a user scrolling through their feed. The user's twitter feed is saved, a predetermined prompt stored in an object is given to a gemini AI to return a bias score. The amount of time a user spends on a post is collected and attributed to an attention statistic along with a few other data points. This data is displayed on the user's screen with data visualization such as a

geiger counter for detecting political bias and a slider scale to represent the percentage of biased posts observed. A GitHub repository exists, link to repository: <https://github.com/spookyuser/intransient> Number of posts viewed per session, pageURL, corpus, bias score, timeInViewPort, session Duration, modelName (gemini version) and time post was recorded. Two people were selected to browse Twitter as they would normally. A total of 9769.008 seconds (162 minutes) over 10 sessions was recorded. Of those sessions 5 were used to gauge the accuracy of the left/right prediction by visiting profiles such as Donald Trump and Bernie Sanders.

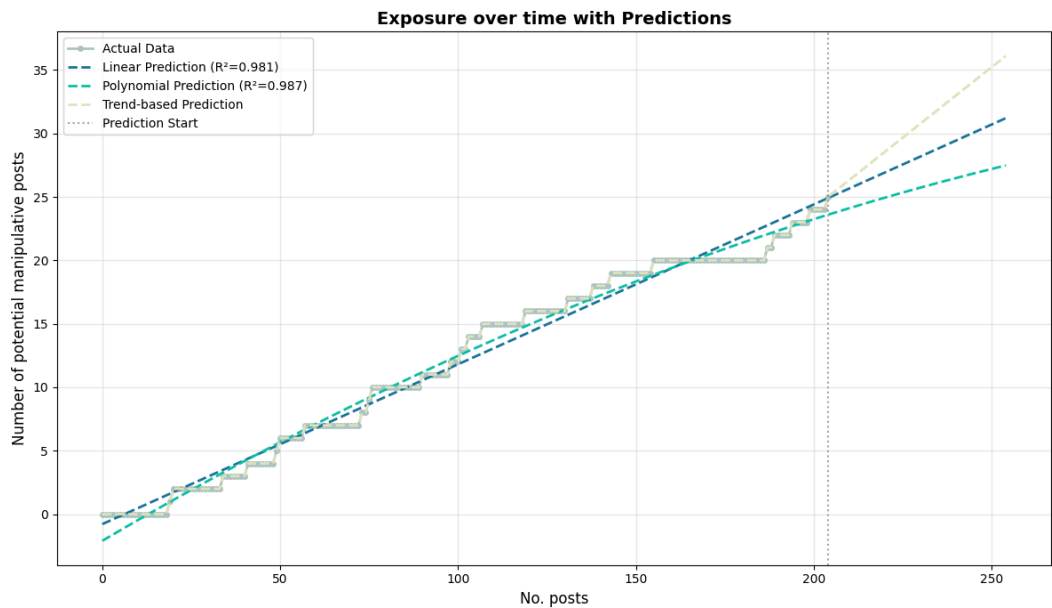
3. Results

Graph showing predicted amount of biased posts with a session time of 7 seconds:

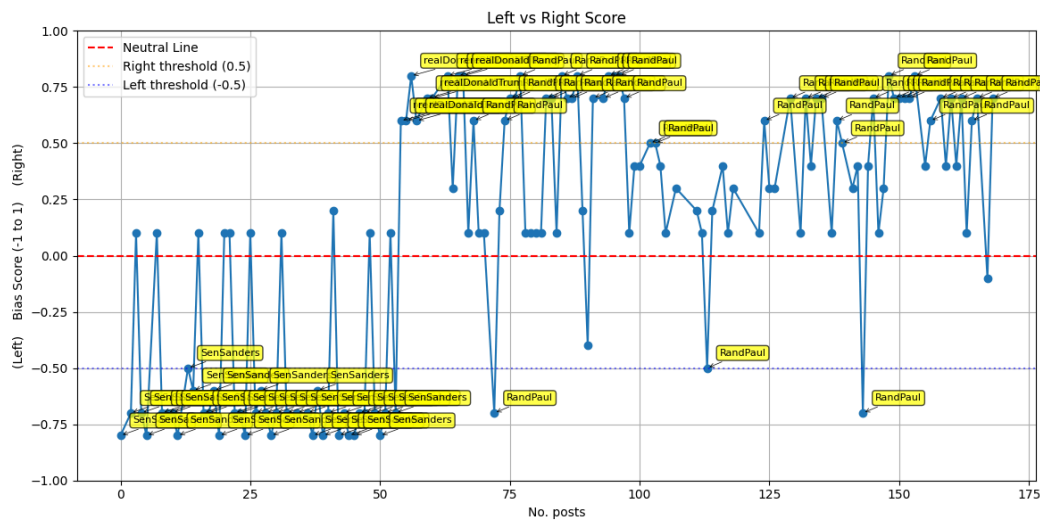


[contd]

Graph showing predicted amount of biased posts with a session time of 1221 seconds:

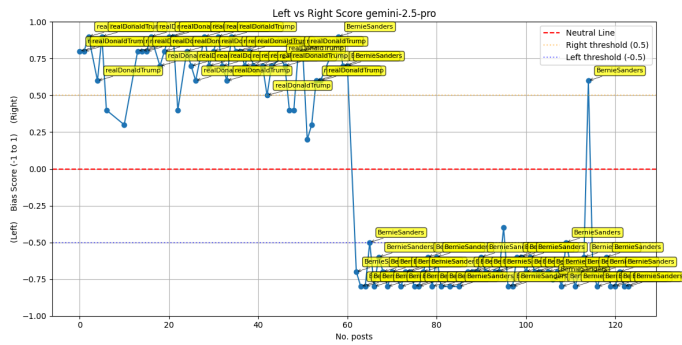


Graph to display where your feed lies (excluding null value bias posts). URLs of posts with a bias over 0.5 or under -0.5 displayed.

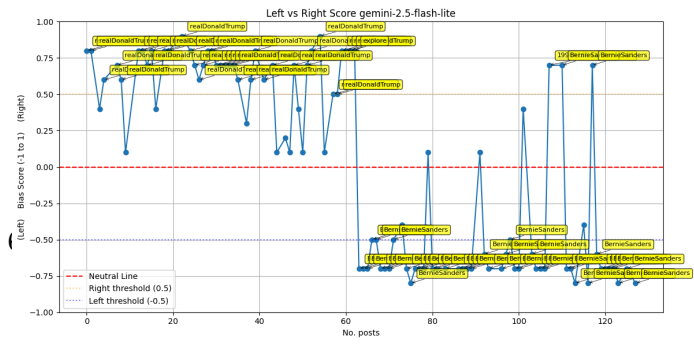


Simulations with three models

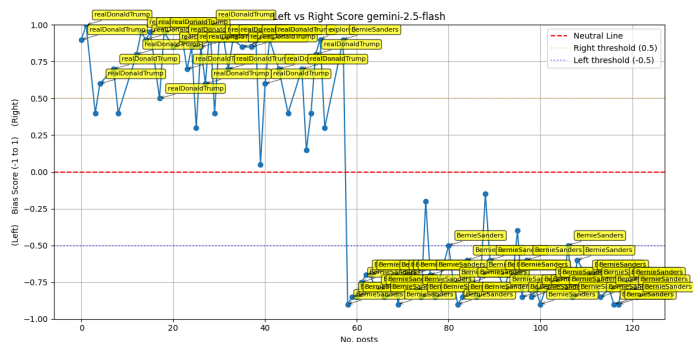
Gemini-2.5 pro



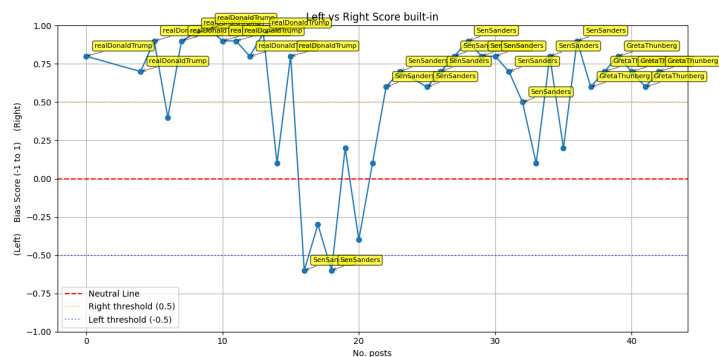
Gemini 2.5-flash-lite



Gemini 2.5-flash



Built-in (Chrome)



4. Discussion and Conclusion

We found that even if you have a small amount of political bias in absolute terms, that will still add up over time. Because of the illusory truth effect, this will have a stronger and stronger impact on your ability to assess content fairly. Certain users like Donal Trump have clear political stances that swing right. Similarly, Bernie Sanders has clear political stances that swing left. It's clear when looking at their profiles with our extension that this is the case. Making users aware of this on its own is likely not that useful. However, on a user's own feed, the extension can reveal a subtle political stance that the user might not be aware of. We believe this additional awareness on more ambiguous areas of twitter, can improve the extent to which users are manipulated by text and therefore by ai generated text. Chrome built-in is not a powerful tool to use for bias classification. Gemini-pro-2.5 is the most accurate engine for labeling posts. Users are exposed to a vast amount of persuasive content.

5. Acknowledgments

Thanks to David G. Rand, Luke Hewitt and Cameron Jones for organizing this Hackathon.

6. References

- [1] Hassan, A. and Barber, S.J. (2021). The Effects of Repetition Frequency on the Illusory Truth Effect. *Cognitive Research: Principles and Implications*, [online] 6. doi:<https://doi.org/10.1186/s41235-021-00301-5>.
- [2] University of Maine (2023). *Social Media Statistics Details*. [online] Undiscovered Maine. Available at: <https://umaine.edu/undiscoveredmaine/small-business/resources/marketing-for-small-business/social-media-tools/social-media-statistics-details/>.

7. Appendix

Limitations, Dual-Use & Security Considerations

Limitations of our approach include scalability - money. External models are expensive to run for multiple users. If we use local models, it can theoretically be free, however in practice it's slow and uses a lot of system resources.

Dual-use risks: Could your method be used to train better manipulators? If you were trying to get an AI to make manipulative content on twitter, this extension could help you create it faster. It could also be used to classify large amounts of text and that could possibly be used to train better manipulators.