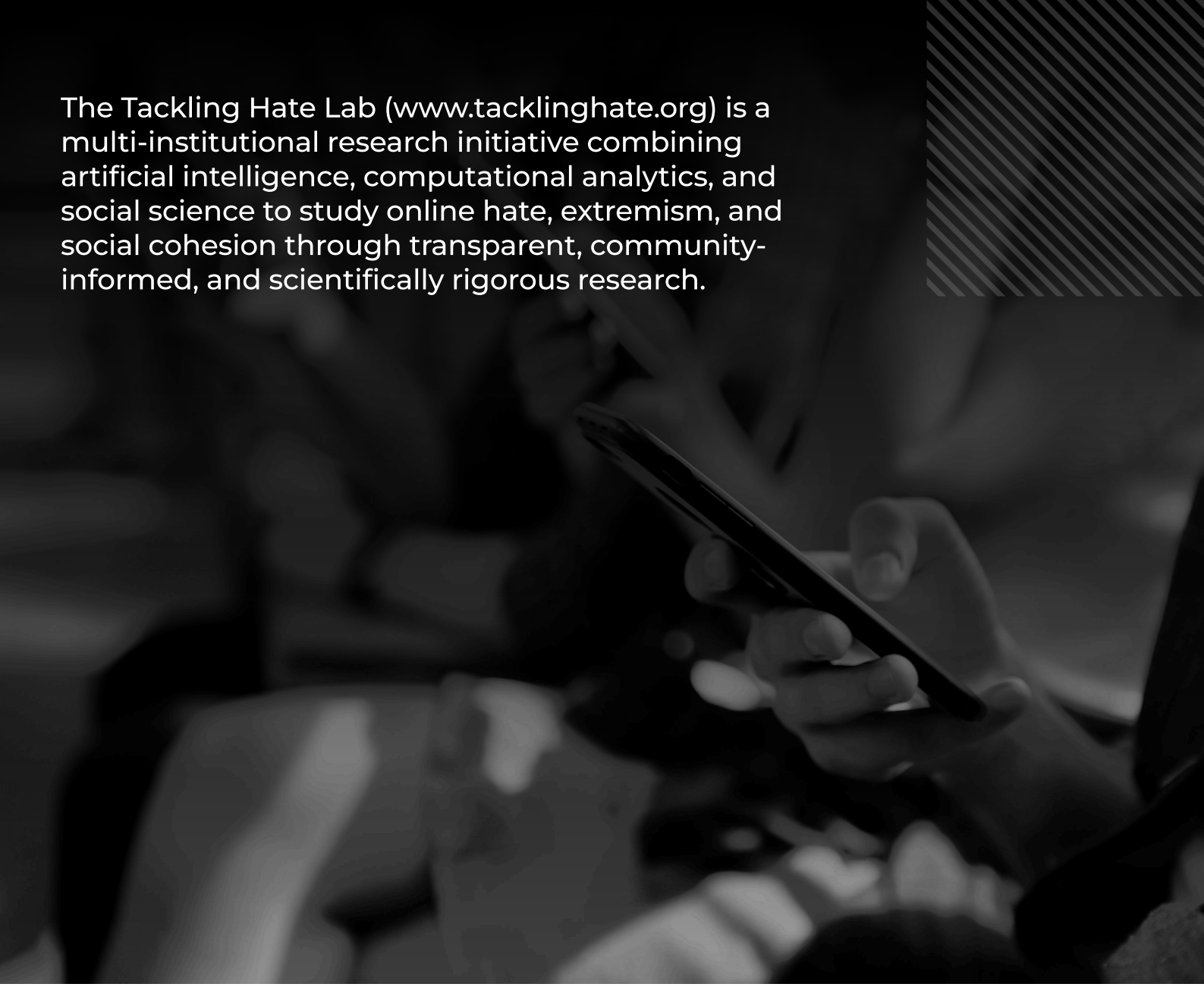
TACKLING HATE

Mapping Online Anti-Jewish Hate in Australia:

A Big Data Analysis

Matteo Vergani, Andrea Giovannetti, Stephanie Ng, Kewen Liao,
Dan Goodhardt, Haily Tran, Huu Phuc Hong, Yinsong Chen

The Tackling Hate Lab (www.tacklinghate.org) is a multi-institutional research initiative combining artificial intelligence, computational analytics, and social science to study online hate, extremism, and social cohesion through transparent, community-informed, and scientifically rigorous research.



Suggested citation: Vergani, M., Giovannetti, A., Ng, S., Liao, K., Goodhardt, D., Tran, H., Hong, H.P., & Chen, Y. (2026). Mapping Online Anti-Jewish hate in Australia: A Big Data Analysis. Tackling Hate Lab.

This report contains discussions of hateful and distressing language, including explicit examples, as well as themes of violence and discrimination. Some content may be confronting or upsetting for readers.

Please engage with this material at your own pace and seek support if needed.

Contents

EXECUTIVE SUMMARY

Aims and scope

This report uses a big data approach to examine online anti-Jewish hate in Australia between October 2021 and March 2026. Drawing on multiple datasets, classification systems, and computational methods, it analyses trends in anti-Jewish, anti-Zionist, and anti-Israel toxicity and identity attack, as well as "old" and "new" antisemitism. The report also examines the relationship between online hostility and real-world incidents targeting Jewish communities, and investigates the role of conspiracy theories in amplifying online hate. By presenting these categories separately and transparently, the report allows readers using different definitional and conceptual frameworks for anti-Jewish hate to interpret trends in a scientifically rigorous and replicable way.

Data

Three datasets underpin the analysis:

- Israel-Gaza dataset: 1,000,306 posts from X (October 2021 – March 2026) on Israel, Palestine, Gaza, and related topics.
- Adass dataset: 173,068 posts from X (December 2024 – January 2025) following the attack on the Adass Israel Synagogue in Melbourne.
- Real-world incidents dataset: 533 anti-Jewish incidents that took place in Australia (October 2022 – April 2025) coded for 7 variables across 5 behaviour types.

Key Findings

1. Analyse online trends.

Online hostility targeting Jews, Zionists, and Israel increased sharply and persistently after 7 October 2023, representing the most significant structural shift observed across the entire study period. Before October 2023, toxicity and identity attack content targeting these groups remained relatively low and sporadic. Following the Hamas terrorist attack and subsequent war in Gaza, **all measures rose substantially, with the strongest increases observed in content targeting Israel and Zionists.** Content directly targeting Jews also increased markedly, particularly identity attack content, indicating a rise in explicit hostility directed at Jewish identity groups. The escalation was not temporary. Although levels declined from their initial peaks during 2024, they remained substantially above pre-October 2023 baselines throughout the remainder of the observation period, suggesting a partial normalisation of hostile discourse online. The report also found that "new" antisemitism, increased at a far greater scale than "old" antisemitism based on traditional anti-Jewish tropes and stereotypes. The Bondi terrorist attack on 14 December 2025 generated a further, smaller short-term increase across most measures.

2. Examine the relationship between online anti-Jewish hate and real-world incidents targeting Jewish communities in Australia.

Physical violence and symbolic aggression against Jewish targets often triggered spikes in online content perceived as anti-Jewish by our community annotators. On average, when a physical act of violence occurred offline, the number of anti-Jewish tweets increased by about 3.4 posts above the baseline level expected over the next 24 hours. One plausible explanation is that such incidents attract media coverage and public attention, which increases their visibility in public discourse and creates opportunities for individuals with anti-Jewish views to express and reinforce those views online. In other words, real-world violence emerges as a stronger driver of online anti-Jewish hate than the reverse.

Our analysis shows that the online and offline spheres have become more tightly interconnected since October 2023. For policymakers, this has serious implications: **monitoring and responding to one without addressing the other risks missing how they reinforce each other.** Interventions must be designed with this feedback loop in mind, combining digital platform governance, community engagement, and early intervention strategies.

3. Investigate the relationship between anti-Jewish conspiracy theories and online hate.

An analysis of one conspiracy theory, which alleged that the Adass Synagogue fire was staged by the Jewish community, found that each tweet promoting this conspiracy triggered, on average, 1.6 anti-Jewish tweets. This conspiracy sustained anti-Jewish hate over long periods and was self-perpetuating. A user-level study comparing the behavioural orbit users with historically very similar levels of online anti-Jewish posting, found that individuals amplifying this conspiracy, increased their anti-Jewish posting by 10.6 percentage points and toxicity by 3.5 points compared with matched controls. This suggests that **users who are exposed to conspiracy-related content are more likely to post anti-Jewish messages,** when other factors are held constant, at least in this context. This pattern is consistent with psychological mechanisms such as confirmation bias and motivated reasoning, whereby individuals preferentially adopt and share narratives that align with their pre-existing beliefs, including anti-Jewish attitudes when conspiratorial claims contain anti-Jewish elements.

RECOMMENDATIONS

Based on the findings of this report, we propose four recommendations aimed at strengthening Australia's capacity to monitor, understand, and respond to online anti-Jewish hate. These recommendations focus on improving early warning and monitoring systems, integrating online hate into broader social cohesion and security frameworks, supporting more effective platform responses, and increasing transparency and research access to platform data.

R1. Fund a National Online Hate Observatory

We recommend that the Department of Home Affairs fund a stable and independent Observatory of Online Hate that uses transparent artificial intelligence methods to monitor anti-Jewish hate and other forms of online hate in near real time. The Observatory should analyse different forms of hostile discourse separately, including anti-Jewish, anti-Zionist, and anti-Israel hostility, as well as old and new antisemitism, to provide rigorous and transparent evidence without imposing a single definitional framework. The findings of this report demonstrate that online hostility increased sharply and persistently after 7 October 2023, with further escalation following the Bondi terrorist attack, highlighting the need for continuous monitoring, early warning capability, and evidence-based policy responses.

R2. Integrate online hate monitoring into social cohesion, community safety, and counterterrorism frameworks

This report demonstrates that online and real-world anti-Jewish hate now operate as a connected system, particularly following October 2023. Physical violence, symbolic aggression, and other real-world incidents were consistently associated with measurable increases in online anti-Jewish hate, suggesting that offline attacks can act as catalysts for broader waves of hostility online. Government agencies responsible for social cohesion, counterterrorism, and community safety should therefore integrate online hate monitoring into existing threat assessment and prevention frameworks. Responses to hate should address the interaction between online and offline environments rather than treating them as separate domains.

R3. Support evidence-based moderation and intervention practices targeting conspiracy amplification

Platforms operating in Australia should be required to detect and respond more effectively to conspiracy narratives linked to offline attacks and community incidents, particularly false flag and staged-event claims. This report found that conspiracy narratives surrounding the Adass synagogue attack played a major role in amplifying and sustaining anti-Jewish hate online, with each conspiracy-related post generating substantially more anti-Jewish content over time. The findings suggest that conspiratorial narratives can function as powerful accelerants of hate ecosystems and should therefore be treated as a priority area for platform intervention, moderation, and risk mitigation.

R4. Improve transparency and independent research access to platform data

Independent researchers should be provided with secure access to platform data to support ongoing monitoring, evaluation, and public accountability regarding online hate trends in Australia. This report demonstrates the value of large-scale computational analysis for identifying structural shifts, trigger-event dynamics, and the interaction between online and offline hate. However, meaningful public-interest research remains constrained by limited access to platform data. Improved transparency and research access are necessary to enable independent evaluation of platform moderation practices, algorithmic amplification, and the spread of harmful narratives during periods of heightened social tension.

The larger issue is identifying when online hate becomes sustained, self-reinforcing, and connected to offline harm. Australia needs monitoring, prevention, and response systems that can detect escalation early and protect democratic debate while reducing identity-based harm.

INTRODUCTION

1.1. Aims

This report uses a big data approach to examine online anti-Jewish hate in Australia between 1 October 2021 and 1 March 2026, drawing on a series of independent studies that apply diverse computational methods, classification systems, and multiple datasets.

The study has three primary aims:

1. Analyse trends in anti-Jewish, anti-Zionist, and anti-Israel toxicity and identity attack, as well as "old" and "new" antisemitism;
2. Examine the relationship between online hostility and real-world incidents targeting Jewish communities in Australia;
3. Investigate the relationship between anti-Jewish conspiracy theories and online hate.

1.2. Data

This report uses three datasets: three capturing online language and one capturing real-world anti-Jewish incidents. The online language datasets were collected using tailored sets of keywords and hashtags to ensure relevance to specific events and topics.

The first online language dataset **(n = 1,000,306)** captures discussions on X related to the Israel–Palestine conflict, collected between 1 October 2021 and 1 March 2026. This extended timeframe enables analysis of both immediate reactions to key events (such as the 7 October 2023 attacks and the 14 December 2025 Bondi terrorist attack) and longer-term trends in conversations about Israel, Palestine, and anti-Jewish hate. The keyword list was designed to reflect a wide spectrum of views, including support for both Israeli and Palestinian communities, reactions to major events, and discussions of Zionism, Judaism, human rights, and military action. The research team selected them using three criteria: relevance to the conflict, frequent use in online conversations about Israel and Palestine, and the inclusion of terms used by different political positions to avoid bias in data collection. The list therefore combines geopolitical references (e.g., #IsraelPalestine, #Gaza, #WestBank), religious or cultural terms (e.g., #Judaism, #Shabbat, #Torah), and advocacy slogans or campaign hashtags (e.g., #IStandWithIsrael, #FreeGaza, #BoycottIsrael).

The second online language dataset **(n = 173,068)** captures X posts focusing on the online response to the attack on the Adass Israel Synagogue in Melbourne on 6 December 2024. It covers the period from 6 December 2024 to 6 January 2025. For the second dataset, the keywords and hashtags used to retrieve the data capture online discussions referring to Jewish identity, religious life, and debates around anti-Jewish hate. The research team selected them using three criteria: discussions about the Adass synagogue fire, relevance to Jewish communities and institutions, public discussions of anti-Jewish hate in Australia (both old and new). The list therefore combines identity and cultural markers (e.g., Jew, Jewish, Ashkenazi, Sephardic), religious or community institutions (e.g., synagogue, shul, shule), ideological terms (e.g., Zionist, Zionism, antisemitism), and location or incident-related references associated with the Adass fire (e.g., #Adass, #Ripponlea, #attack).

The two datasets were constructed using different keyword parameters, resulting in different daily tweet volumes. The second dataset records a higher average daily volume and contains discussions more directly focused on anti-Jewish hate. In contrast, the first dataset spans a longer period and captures a lower daily volume of tweets overall. The key difference between the keyword parameters is that the first dataset used broad conflict-related terms associated with the Israel–Palestine debate, allowing us to capture a wide spectrum of political and public discourse. In contrast, the second dataset used more targeted keywords related to Jewish identity, anti-Jewish hate, and the Adass synagogue attack, enabling us to isolate conversations specifically focused on the incident and its interpretation. Given these differences in search criteria, thematic focus, and temporal coverage, the total volumes of anti-Jewish hate-related content across the two datasets are not directly comparable.

The rationale for using X data in the first two datasets is its unique capacity to support large-scale, real-time analysis of online discourse, despite being used by a relatively small proportion of Australians (around 17%, or 4.7 million users). While not fully representative of the population, this user base captures a highly active segment of public conversation. X offers a rare combination of scale, openness, and accessibility, allowing researchers to collect posts alongside engagement metrics such as replies, reposts, and likes. Importantly, geolocation metadata enables the identification of Australian-based content, ensuring analytical relevance. In contrast, platforms like Facebook and Instagram restrict large-scale data access. Since its acquisition by Elon Musk in 2022, X has also become a key site for both mainstream and fringe discourse, making it particularly valuable for tracking the emergence and spread of hate narratives in real time.

The third dataset captures **533** real-world anti-Jewish incidents that occurred in Australia between October 2022 and April 2025. Data sources included mainstream media, police media releases, community organisations including Executive Council of Australian Jewry (ECAJ), Anti-Defamation Commission (ADC), Jewish Community Council of Victoria (JCCV), institutional statements, and social media posts referring to offline events. Incidents were included if they occurred in physical or real-world settings within Australia and involved anti-Jewish content. Where incidents in ECAJ, ADC or JCCV reports required clarification or fell outside the most recent reporting period, the research team supplemented records with additional publicly available sources to improve contextual detail and ensure consistent classification across the study period. To ensure consistency, only incidents with a clearly identifiable date or month were included. Incidents without a clear timeframe were excluded. Repeated actions forming part of a single coordinated episode, such as recurring graffiti within a short period, were coded as one incident. Events involving multiple forms of anti-Jewish behaviour, such as vandalism and verbal abuse, were also recorded as a single incident with multiple behavioural categories. Online-only incidents were excluded. Each incident was manually reviewed and coded independently by three researchers using a structured coding framework. Variables included date, location, behaviour type, severity, number of victims, police involvement, media coverage, and perpetrator characteristics where available. Ambiguous cases were resolved collaboratively using a shared coding rubric.

1.3. How we measured online anti-Jewish hate in this report

The report separates categories of anti-Jewish, anti-Zionist, and anti-Israel toxicity and identity attack analytically, as well as "old" and "new" antisemitism, using multiple classification systems that allow readers to examine trends in each type of discourse separately. Depending on the definitional framework adopted, some readers may consider all these categories part of a broader superordinate category of anti-Jewish hate, while other readers may view only some categories as anti-Jewish hate and regard others as adjacent or distinct concepts. We therefore present the data in a transparent, scientifically rigorous, and replicable way that allows readers using different conceptual frameworks to define anti-Jewish hate and interpret trends independently.

First, we used Perspective API, a machine learning system developed by Jigsaw to evaluate the likelihood that text will be perceived as harmful or disruptive in online discussions. The tool generates probabilistic scores for different attributes of language. In this report, we focus on two attributes: toxicity, defined as "speech that is rude, disrespectful, or likely to disrupt constructive discussion", and identity attack, which captures language that is insulting, demeaning, or dehumanising towards a person or group based on a protected characteristic, such as religion or ethnicity. While toxicity measures general incivility, identity attack identifies more targeted forms of abuse, including anti-Jewish expressions. Together, these metrics enable a systematic assessment of both broad and group-directed harmful language in online discourse.

Once this classification was applied across the dataset, we used Qwen, an advanced artificial intelligence system trained to interpret human language, to identify who or what each piece of harmful content was directed at. The model was applied consistently across all posts to determine whether the content targeted Jews, Israel, or Zionists. Each target group was coded separately, allowing us to distinguish between different forms of targeting and report them clearly in the results.

Second, we used a machine learning classifier developed by the Tackling Hate Lab, designed to identify anti-Jewish content based on how it is perceived by members of the Australian Jewish community. The model was trained using data annotated by contributors recruited through the Executive Council of Australian Jewry and the New South Wales Jewish Board of Deputies. As a result, the classifier reflects community-informed understandings of anti-Jewish hate, though it does not capture the full diversity of views across all Jewish communities. The model analyses text and classifies whether a post is anti-Jewish or not. It performed strongly across standard evaluation measures. Overall, it was correct 87.48% of the time. When it identified a post as anti-Jewish, it was correct in nearly 88% of cases, indicating a relatively low rate of false positives. It also successfully detected around 85% of all anti-Jewish posts, meaning relatively few were missed. A combined performance score (F1) of 86.35% shows a good balance between these strengths. The model also demonstrated a high ability to distinguish between anti-Jewish and non-anti-Jewish content across different thresholds (AUC: 93.77%), and maintained strong performance when tested on new, unseen data (AUC: 90.59%).

We then used topic modelling to distinguish between different forms of anti-Jewish hate within the content identified by our classifier. Using BERTopic, we grouped similar posts into clusters based on shared language patterns, and applied GPT-4o to generate descriptive labels for each topic. To simplify interpretation, these topics were grouped into "old" and "new" antisemitism. Old antisemitism refers to traditional forms targeting Jews as a religious or ethnic group, including conspiracy myths, Holocaust denial, and incitement to violence. New antisemitism refers to hostility expressed through discourse about Israel, including denying Jewish self-determination or holding Jews collectively responsible for the actions of the Israeli government.

Examples

Anti-Israel toxicity: The IDF is the evil on Earth. It likes and adores murdering children and babies #IsraelTerroristEntity

Anti-Israel identity attack: Fuck Israel

Anti-Zionist toxicity: Zionism is pure terrorism

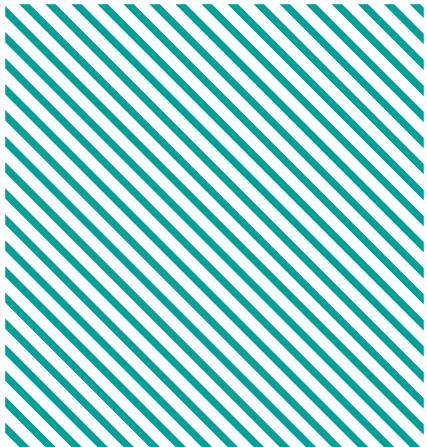
Anti-Zionist identity attack: Its for blood sacrifice by pedo Zionist goblins. They worship Baal

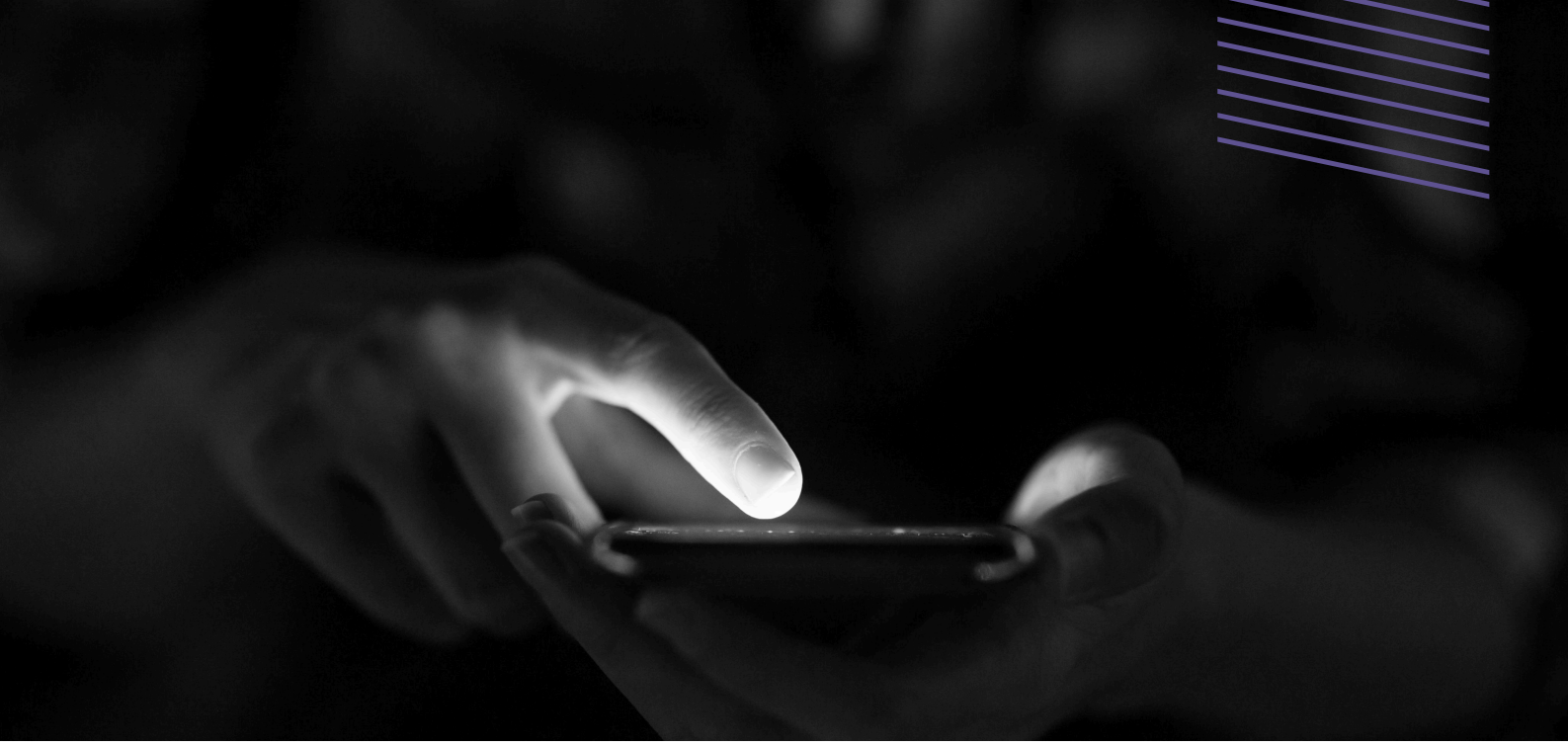
Anti-Jewish toxicity: You fucking Jew

Anti-Jewish identity attack: Throw that fat Jew in the sea. See how well he can swim.

New antisemitism: Israel was founded on terrorism & ethnic cleansing #destroyIsrael

Old antisemitism: I hate Jews





2. TRACKING ONLINE TRENDS

2.1. Tracking trends in anti-Jewish, anti-Israel and anti-Zionist toxicity and identity attack

Figures 1 to 6 show the evolution of targeted toxicity and identity attack content (classified using Perspective API) between October 2021 and March 2026 using 7 day rolling averages. Across all six figures, the most important structural break occurs immediately after 7 October 2023, marked by a sudden and sustained escalation in content containing toxicity and identity attack targeting Jews, Zionists, and Israel. Before this date, all series remained relatively low and sporadic, with most daily averages close to zero and only occasional short-lived spikes linked to isolated events.

The strongest increases occur in content targeting Israel and Zionists. Figures 1 and 2 show that both toxicity and identity attack targeting Israel increased sharply immediately after 7 October 2023, reaching sustained levels above 100 daily posts during October and November 2023, with peaks exceeding 200 daily posts. Figures 3 and 4 show a similar pattern for content targeting Zionists, although at lower overall volumes. Identity attack targeting Zionists reached particularly high levels during late 2023 and early 2024, with several sharp spikes during May and June 2024. These patterns indicate that the post-7 October escalation was not limited to direct hostility toward Jewish identities, but extended strongly to political and symbolic targets associated with Israel and Zionism.

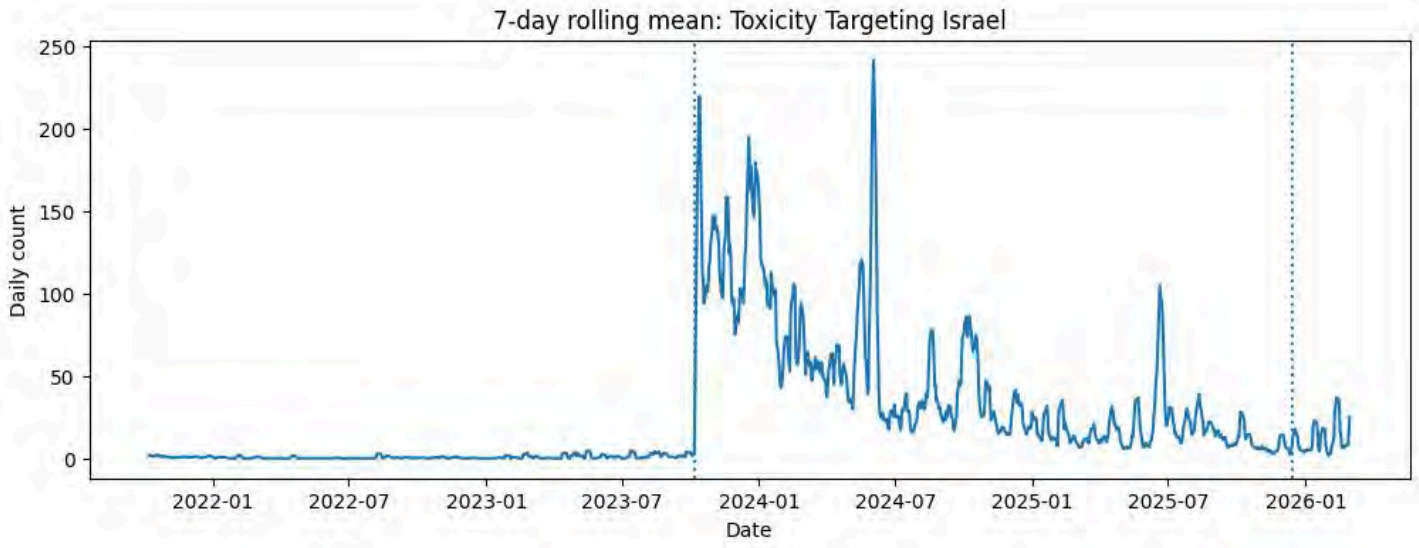


Figure 1. Seven day rolling mean of toxicity targeting Israel, October 2021 to March 2026. Vertical lines indicate 7 October 2023 and 14 December 2025.

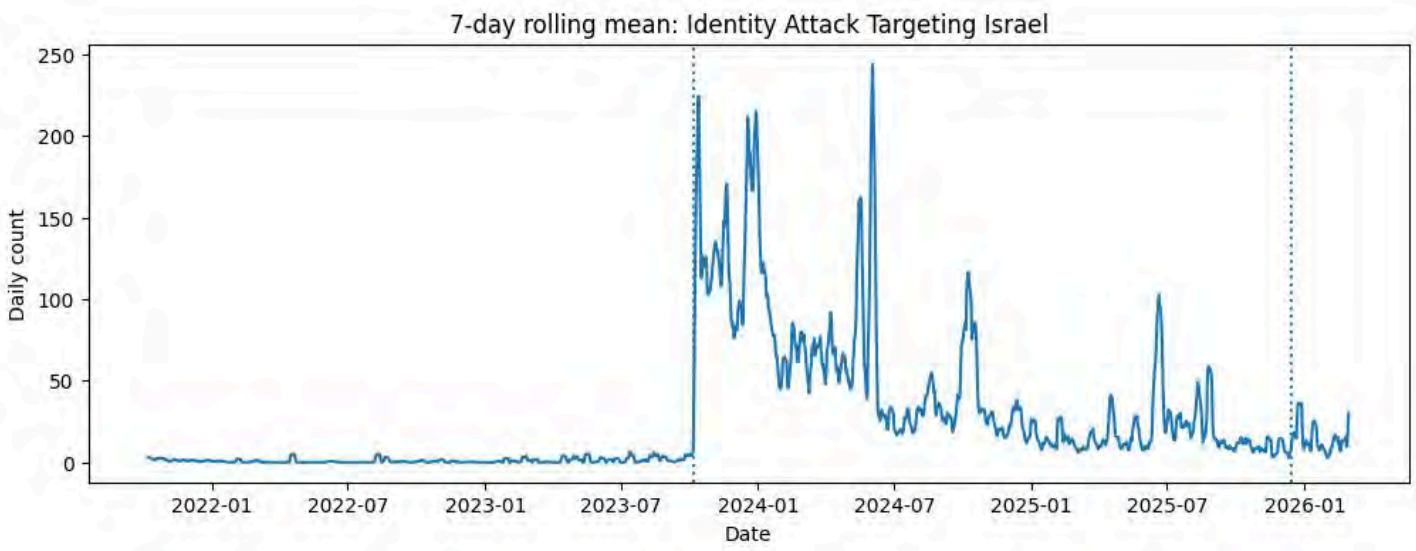


Figure 2. Seven day rolling mean of identity attack targeting Israel, October 2021 to March 2026. Vertical lines indicate 7 October 2023 and 14 December 2025.

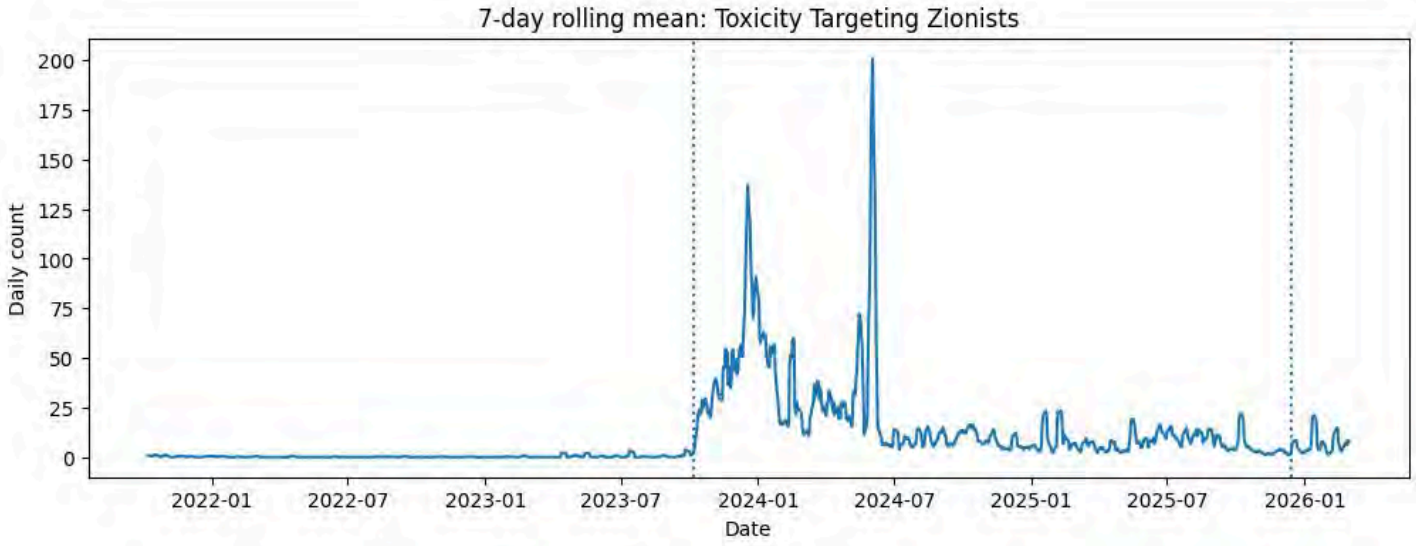


Figure 3. Seven day rolling mean of toxicity targeting Zionists, October 2021 to March 2026. Vertical lines indicate 7 October 2023 and 14 December 2025.

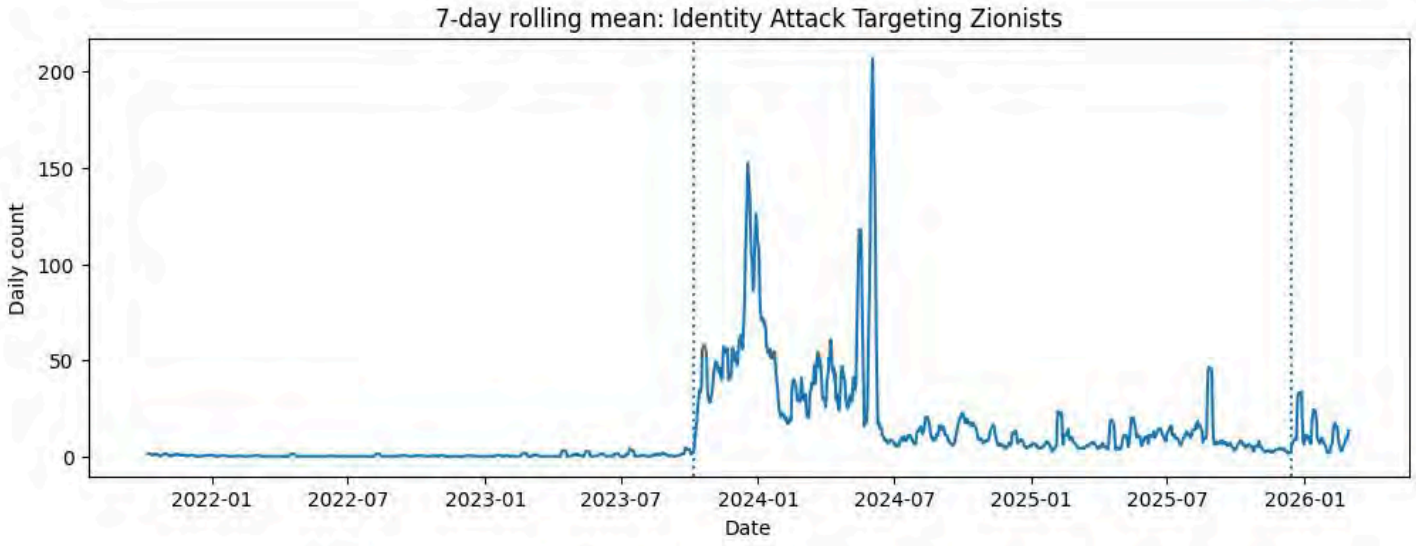


Figure 4. Seven day rolling mean of identity attack targeting Zionists, October 2021 to March 2026. Vertical lines indicate 7 October 2023 and 14 December 2025.

Figures 5 and 6 show that content directly targeting Jews also increased substantially after 7 October 2023, although at lower absolute volumes than Israel-targeted content. Identity attack targeting Jews increased more sharply than toxicity targeting Jews, suggesting that the escalation involved more explicit attacks directed at group identity rather than only generally toxic language. The immediate post-attack period in October 2023 produced the highest sustained levels observed in the series, particularly during the first three months after the Hamas attack and subsequent war escalation.

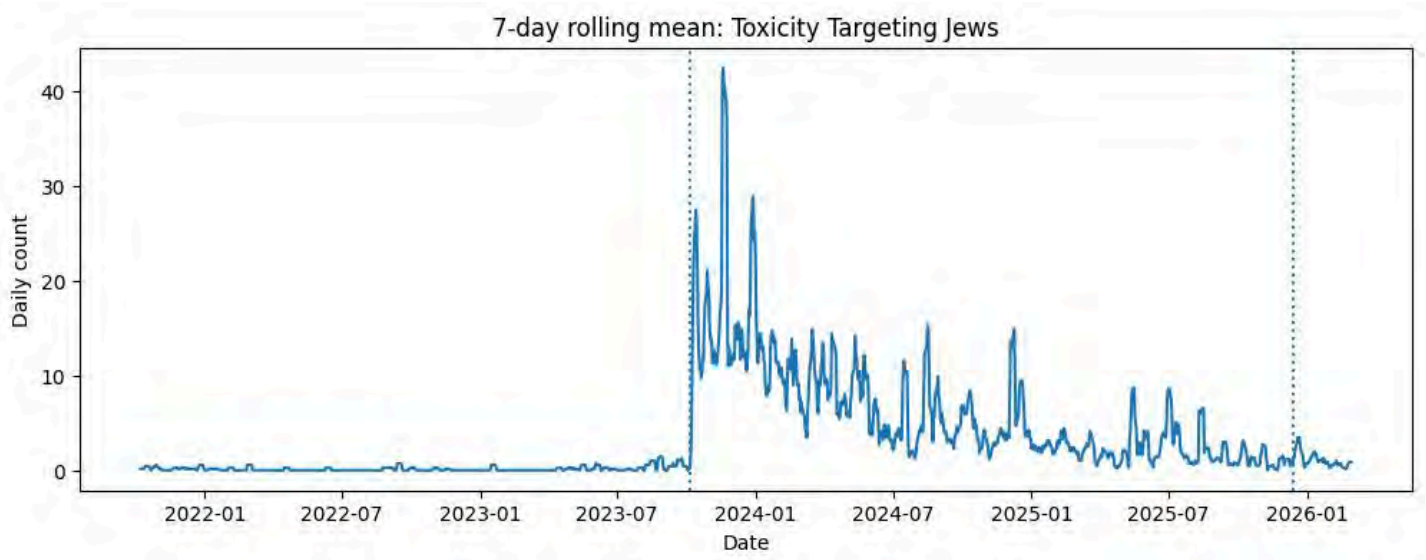


Figure 5. Seven day rolling mean of toxicity targeting Jews, October 2021 to March 2026. Vertical lines indicate 7 October 2023 and 14 December 2025.

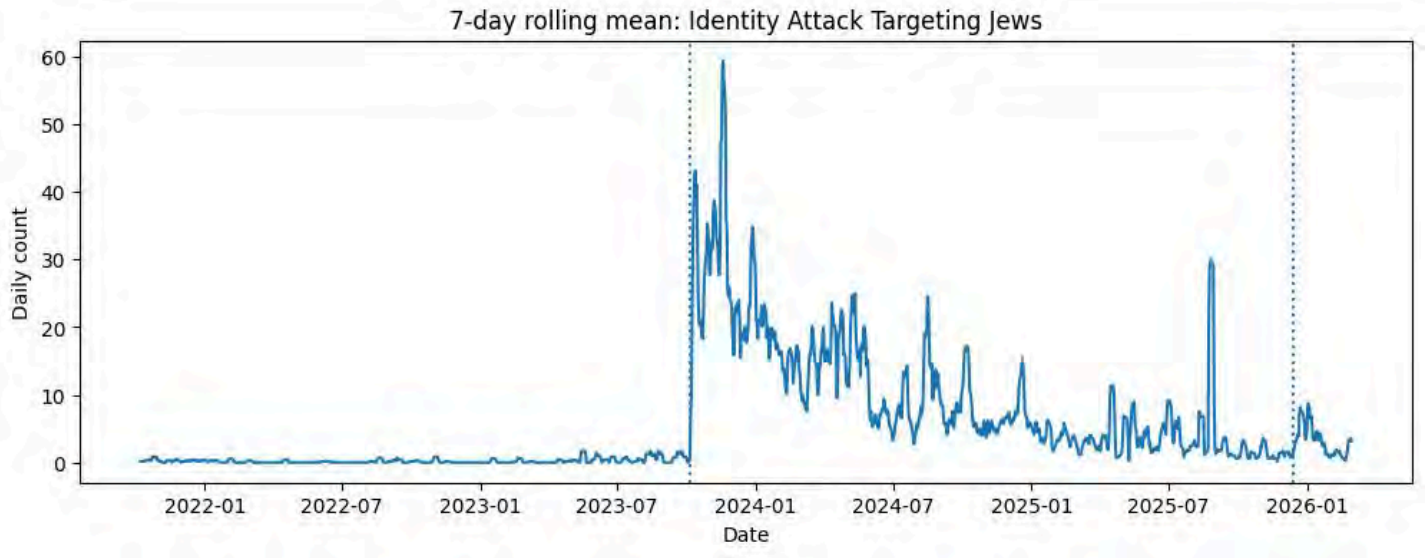


Figure 6. Seven day rolling mean of identity attack targeting Jews, October 2021 to March 2026. Vertical lines indicate 7 October 2023 and 14 December 2025.

Across all figures, the post-October 2023 period follows a similar trajectory. The initial escalation phase between October 2023 and early 2024 is characterised by extremely high volatility, repeated spikes, and sustained elevated baselines. A second period emerges during mid-2024, where overall levels decline but remain substantially above pre-October 2023 levels. Even during quieter periods in late 2024 and 2025, daily averages rarely return to the near-zero baseline that characterised most of 2021 to mid-2023.

This suggests persistence and partial normalisation of hostile discourse rather than a temporary shock. The second marked disruption occurs after 14 December 2025, the date of the Bondi terrorist attack, indicated by the vertical reference line in all figures. Compared with 7 October 2023, this event generated a smaller but still visible increase, particularly in identity attack measures. The increase is clearest in Figures 2, 4, and 6, where short-term spikes appear immediately after the event.

The effect is weaker for general toxicity than for identity attack, indicating that the post-Bondi escalation involved more direct targeting of identity groups rather than broader toxic discussion. However, unlike the sustained escalation observed after October 2023, the December 2025 increase appears shorter and more localised in duration.

Overall, the figures indicate that 7 October 2023 fundamentally transformed the scale and persistence of anti-Jewish, anti-Zionist, and anti-Israel hostile discourse in Australia. The data further suggest that subsequent real-world events, including the Bondi attack period, occurred within an already elevated and destabilised online environment.

2.2. Tracking trends in old and new antisemitism

Figures 7 and 8 show markedly different trajectories for old and new antisemitism, although both exhibit a substantial structural break immediately after 7 October 2023. Prior to this date, both categories remained comparatively low and stable across the observation period. However, the magnitude, persistence, and volatility of the post-October escalation differed substantially between the two forms.

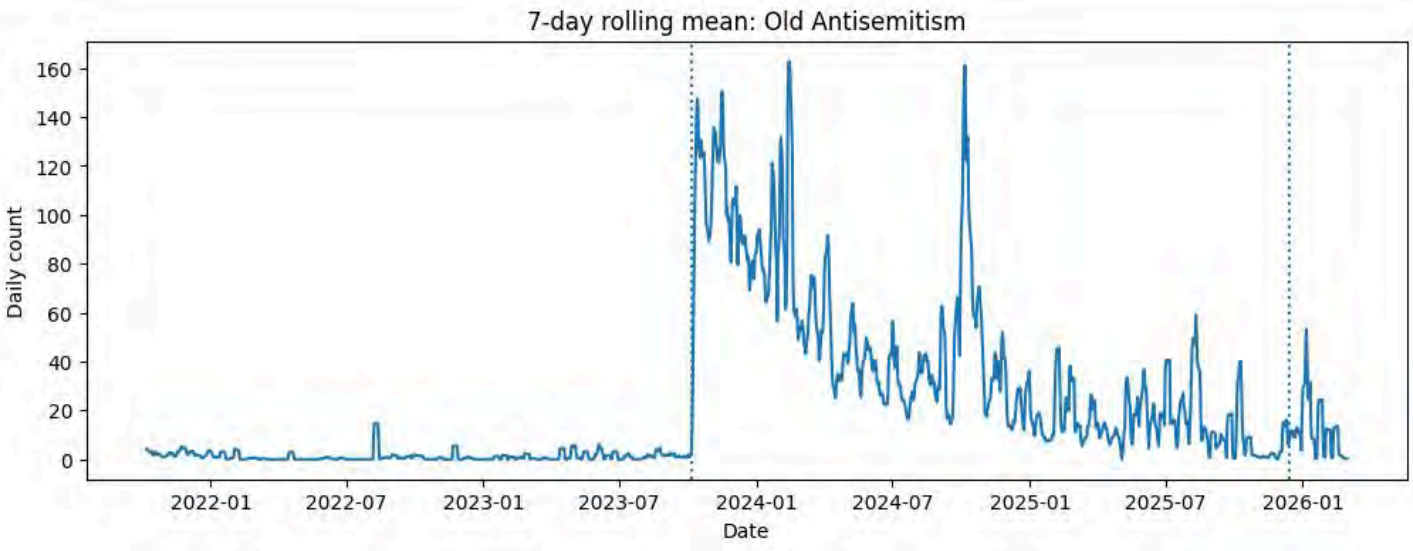


Figure 7. Seven day rolling mean of old antisemitism daily counts between October 2021 and early 2026. Vertical dotted lines indicate 7 October 2023 and 14 December 2025.

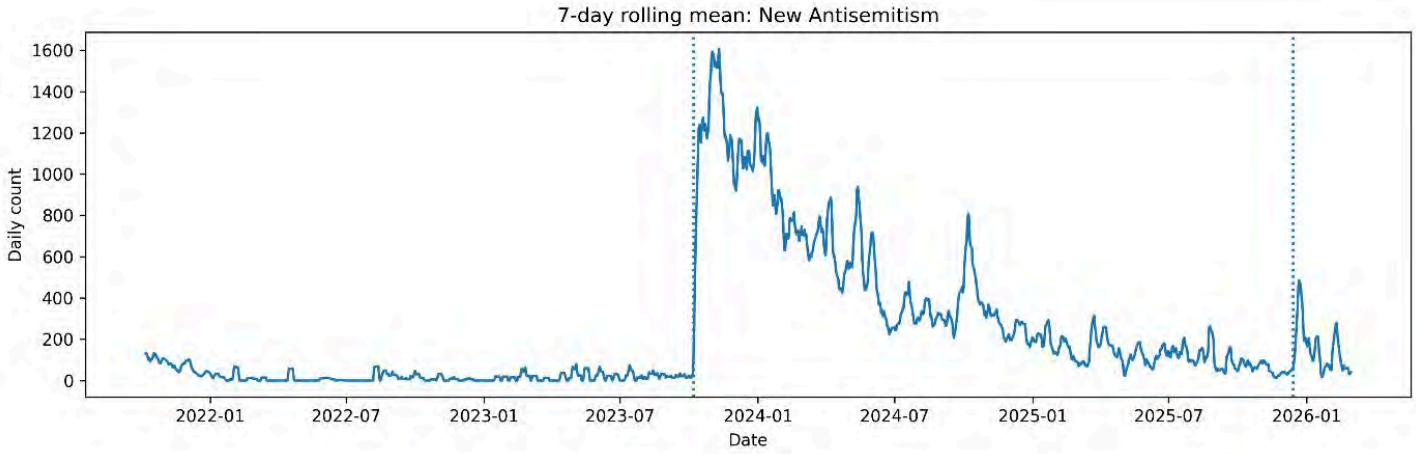


Figure 8. Seven day rolling mean of new antisemitism daily counts between October 2021 and early 2026. Vertical dotted lines indicate 7 October 2023 and 14 December 2025.

New antisemitism shows the most dramatic increase (Figure 8). Immediately after 7 October 2023, the series shifts from a relatively modest baseline to an unprecedented and sustained surge. The 7 day rolling mean rapidly exceeds 1,000 daily posts and peaks at approximately 1,600 posts per day in late 2023. This represents not simply a temporary spike but a prolonged elevated phase lasting throughout much of 2024. Although levels gradually decline over time, they remain substantially above pre-October baselines across the entire post-event period. The pattern suggests that geopolitical developments associated with the Israel Hamas conflict generated a large and sustained expansion of discourse blaming Jews for the actions of Israel or suggesting Israel should not exist and should be eliminated, in ways perceived as anti-Jewish by community annotators, and classified as new antisemitism in this analysis.

Old antisemitism also increased sharply after 7 October 2023, but at a much lower scale (Figure 7). The series rises from a low and stable baseline to repeated peaks between approximately 80 and 160 daily posts during late 2023 and early 2024. Unlike new antisemitism, however, the overall magnitude is considerably smaller, and the decline toward lower levels occurs more quickly across 2024. Nevertheless, levels remain persistently above the pre-October baseline throughout the remainder of the series, indicating that traditional antisemitic narratives and stereotypes also intensified substantially following the conflict.

A further difference concerns volatility. New antisemitism displays large recurrent surges throughout 2024, including major peaks in mid and late 2024, suggesting that this form of hostility was highly reactive to ongoing international developments and media cycles linked to the conflict. Old antisemitism, by contrast, shows a more gradual decay pattern with intermittent spikes, indicating less event driven volatility and a more stable background presence once elevated.

The second marked event, 14 December 2025, is associated with a comparatively modest but visible increase in both categories. In both figures, a short term rise is visible immediately after this date, but the magnitude is substantially smaller than the post-7 October escalation.

3. THE RELATIONSHIP BETWEEN ONLINE LANGUAGE AND REAL-WORLD EVENTS

To examine how trends in online language correspond to real-world incidents in Australia, we cross-analysed language from a subset of the Israel–Palestine conflict dataset (between October 2022 and April 2025) with the ad hoc dataset capturing anti-Jewish incidents that occurred in Australia (between October 2022 and April 2025).

3.1. Statistical modelling

To examine the relationship between online activity and real-world events, we applied a statistical modelling approach known as a Hawkes process. This method enables us to assess whether one type of event, for example the vandalism of a synagogue, can trigger another, such as a surge in social media posts targeting Jewish people, or the reverse. In practical terms, the model helps us identify whether language classified as anti-Jewish by our community members tends to escalate after certain real-world events.

We conducted this analysis in two main stages. In the first, we looked at the relationship between five categories of real-world anti-Jewish incidents captured in our database, which range from verbal abuse to physical violence, and the overall volume of anti-Jewish tweets in our dataset, allowing the model to identify self and cross-effects across the 6 data streams (an online stream and five offline streams). This model covered a wide time window, from 6 October 2022 to 27 April 2025, allowing us to identify patterns well before and after the events of October 7 2023.

In the second stage, we drilled down further. We broke anti-Jewish tweets into the twelve subcategories in Table 2 (such as Holocaust denial, conspiratorial accusations, and anti-Zionism denying Jewish self-determination) and analysed the relationship within and between the streams of real-world incidents and online narratives. To make this comparison meaningful, we divided the timeline into two key periods:

- The Pre-October 7 Period, from 6 October 2022 to 6 October 2023.

- The Post-October 7 Period, from 8 October 2023 to 8 October 2024.

In both exercises, we treated each online and offline stream of events as its own "timeline" within a multidimensional system. The Hawkes model then estimated how much activity in one timeline (e.g. a physical attack on a synagogue) influenced activity in another (e.g. a spike in anti-Jewish posts on X), or even its own (e.g. one anti-Jewish tweet triggering more of the same).

Importantly, we used a non-parametric version of the model, which means we did not impose strong assumptions about how influence operates over time. Instead, we let the data speak for itself by measuring how often one type of event occurred shortly after another. We assumed that any given tweet could influence other tweets for up to 10 hours (with an effect which is assumed to fade with time passing), and we broke this time into 50 equal intervals to track how influence rises and fades.

3.2. What we found

Our first model (in Figure 9) reveals that real-world anti-Jewish incidents do influence online anti-Jewish hate, but not all incidents have the same impact. In particular, we observed sharp increases in anti-Jewish tweets following episodes of physical violence, with an average increment of 3.39 anti-Jewish tweets against the baseline following a single episode. This effect is substantially larger than the coefficients observed for other classes of incidents in the model, indicating that episodes of physical violence are associated with the strongest increases in online anti-Jewish activity. This result suggests that even isolated real-world attacks can act as catalysts for online hate, highlighting a potentially dangerous feedback loop: even though such incidents are relatively rare, they appear to act as powerful triggers for spikes in online hate.

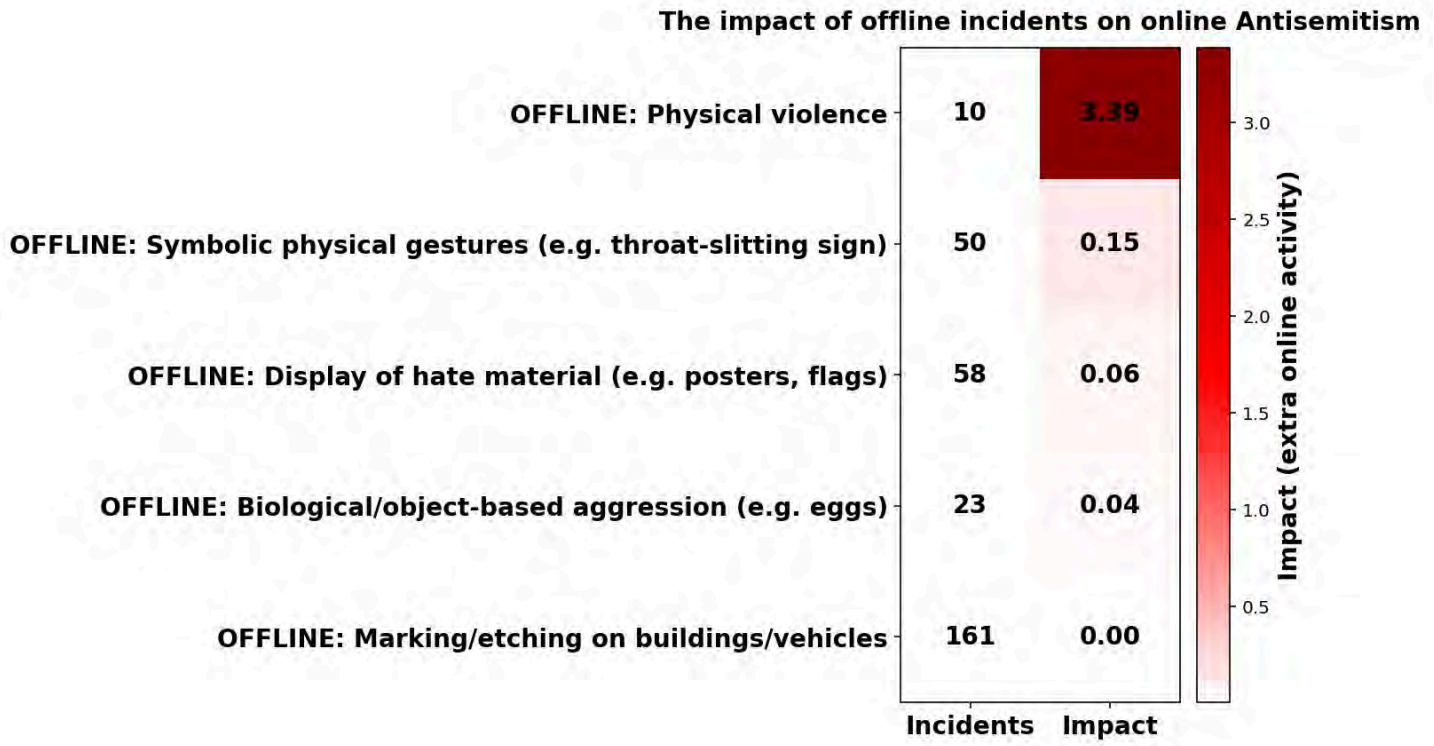


Figure 9. Estimated influence (as measured in terms of extra online activity) of real-world anti-Jewish incidents on online anti-Jewish activity

Subsequently, when we compared the Pre- and Post-October 7 periods, the contrast was stark. Figures 10 and 11 show results from models based on two different time periods, before and after the October 7 attacks. Each figure is a matrix that shows how different types of real-world events and different types of online discourse influence one another. In these matrices, the rows and columns represent different event and discourse types. If a cell is coloured (from white to black), it means there is an effect: the realisation of an event of the type captured in the column increases the probability of events associated with the event type in the row.

Before October 7 (Figure 10), each type of event mostly influenced itself (an effect shown along the diagonal) and there was little interaction between different narratives or between online and offline events. In other words, most activity followed a pattern of self-excitation, where individual narratives, whether online or offline, tended to circulate within their own space. Cross-influence between online and offline events was limited, and the discourse remained relatively fragmented.

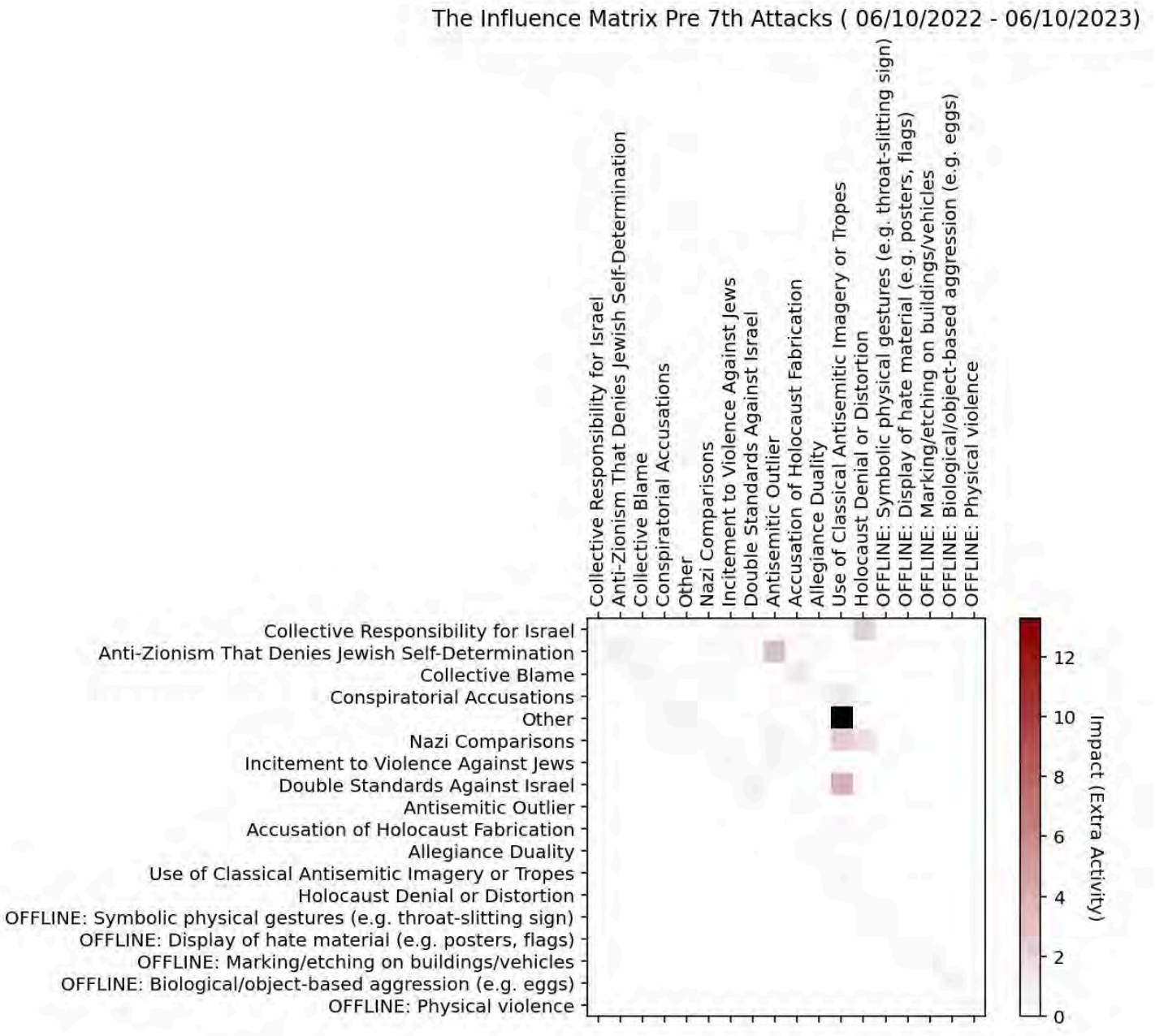


Figure 10. Excitation effects between real-world incidents and online anti-Jewish narratives (Pre-October 7 period).

After the attack (Figure 11), the picture changes significantly. There is much more cross-influence between different event types, especially between real-world incidents and online narratives. This means that real-world events were far more likely to trigger online reactions, and vice versa. In other words, what happens in the real world (such as a physical assault or the vandalism of a synagogue) now has a much stronger and faster impact on what people say and share online. We see that some online narratives are now more likely to be triggered by other narratives or real-world events, suggesting a much more interconnected and reactive environment.

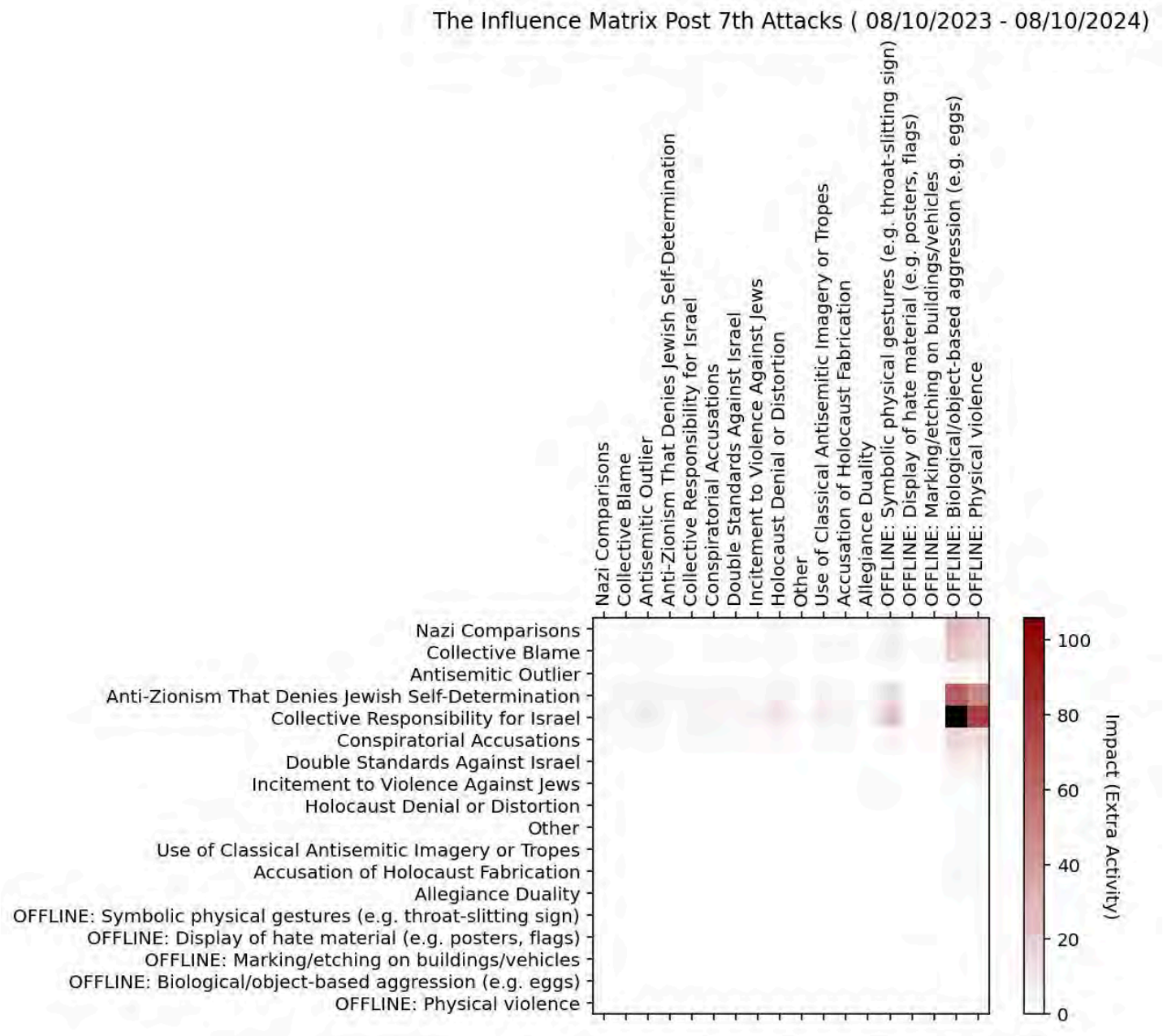


Figure 11. Excitation effects between real-world incidents and online anti-Jewish narratives (Post-October 7 period).

Notably, three of the five categories of real-world incidents were consistently followed by increases in specific types of online anti-Jewish content only after October 7, not before. These categories include symbolic physical gestures (such as the Nazi salute or the throat-slitting sign), object-based aggression (like throwing eggs), and acts of physical violence. Each of these incident types was linked to spikes in particular forms of anti-Jewish rhetoric online, especially content that frames all Jews as collectively responsible for the actions of Israel, denies Jewish rights to self-determination, attributes collective blame to Jews, or makes comparisons between Jews and Nazis.

We hypothesise that the reason why real-world anti-Jewish incidents post-October 7 are associated with spikes in online anti-Jewish hate (both old and new) is that these incidents are more likely to receive media coverage and public attention. This increased visibility brings anti-Jewish hate into mainstream discourse, creating openings for individuals with anti-Jewish views to express and reinforce their hatred online. In this context, media attention does not merely reflect real-world events but may also amplify their impact by providing moments of heightened visibility that anti-Jewish actors exploit to spread harmful narratives and escalate online hate.

4. ANTI-JEWISH CONSPIRACY THEORIES

To investigate the role of conspiracy theories in amplifying and sustaining online hate, we investigated the dynamics between conspiratorial and anti-Jewish content on social media in the wake of the Adass Israel Synagogue attack in Melbourne on 6 December 2024.

We used a dataset capturing the online response to the attack, covering the period from 6 December 2024 to 6 January 2025 and containing 173,068 posts, all sourced from X. Within this dataset, we identified 14,337 users talking about the event in Australia.

Each tweet in the dataset was classified using our community-informed classifier capturing anti-Jewish hate. Additionally, we trained a model to identify tweets spreading the conspiracy that the Adass attack was self-inflicted by the Jewish community to conspiratorial claims that the attack may have been a "false flag" perpetrated by Zionists for them to distract from Israel action in Gaza.

The central aim of this study was to understand whether conspiracy theories drive anti-Jewish hate online, whether the reverse is true, or whether the relationship is mutual.

Initial descriptive analysis revealed a strong correlation between anti-Jewish and conspiratorial narratives, both at the hourly and daily levels (see Figure 12). This relationship was further visualised through a scatterplot of both time series (Figure 13), confirming the high degree of association between them. However, while these correlations are suggestive, they do not reveal directionality, that is, whether one type of content is driving the other or if both are driven by external factors.

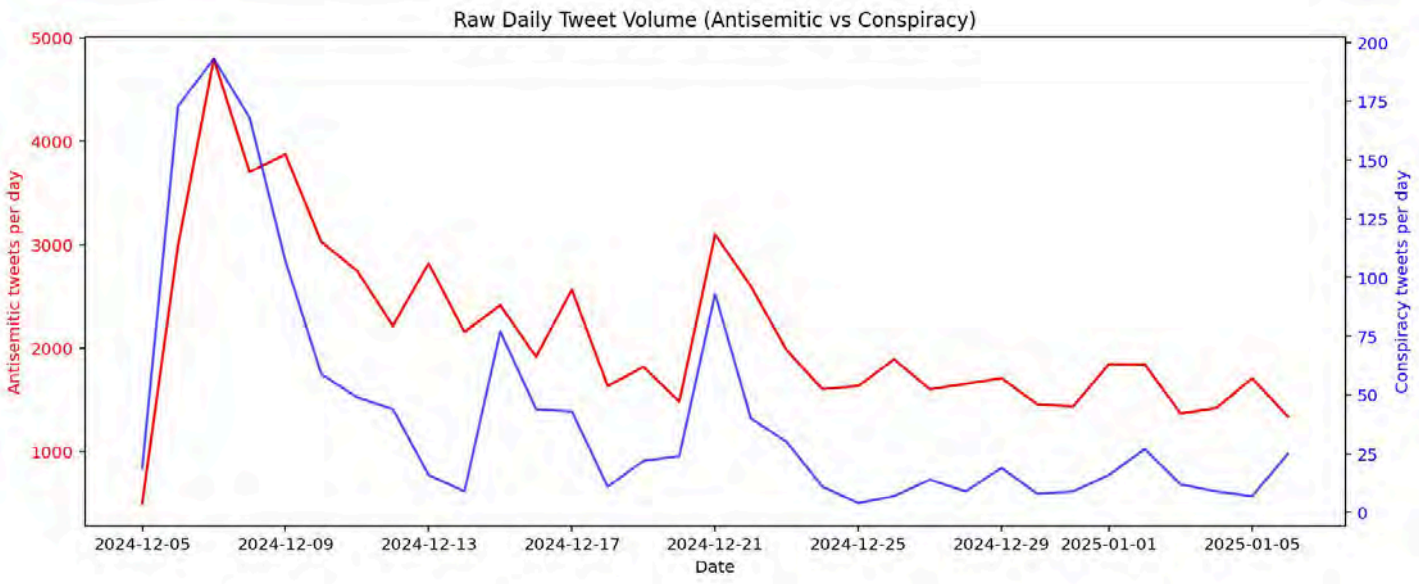


Figure 12. Daily levels of anti-Jewish content and content spreading the conspiracy that the Adass attack was self-inflicted by the Jewish community.

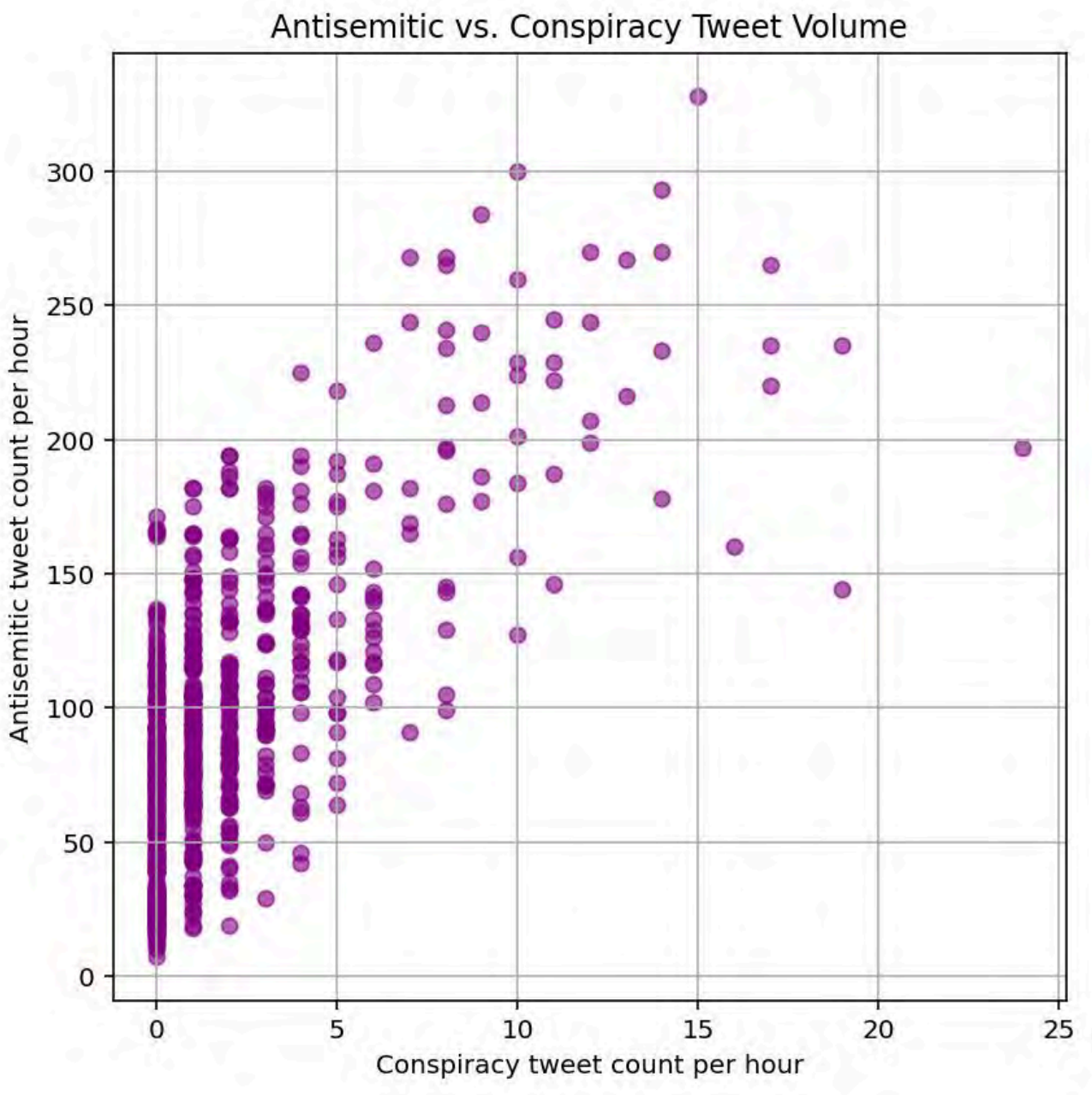


Figure 13. Scatterplot of the time series of anti-Jewish hate and conspiracy content.

To investigate causality, we modelled the dynamics of these narratives using a multidimensional Hawkes process. This type of model is specifically designed to estimate whether one stream of events (e.g., conspiracy tweets) triggers another stream (e.g., anti-Jewish tweets), or whether both are primarily self-exciting, meaning they evolve based on their own internal momentum.

Given no strong prior assumptions about the structure of the influence, we adopted a non-parametric estimation approach using a conditional laws estimator. This method estimates the likelihood that one event triggers another by computing empirical averages based on observed event sequences.

To implement the model, we transformed the streams of anti-Jewish and conspiratorial tweets into chronologically ordered event sequences, using a maximum influence window of 24 hours. This 24-hour period was divided into 50 evenly spaced time intervals, allowing us to estimate how the triggering influence of a tweet decays over time with hourly resolution.

First, we confirmed that the process was statistically stable. The spectral radius of the kernel norm matrix—used to determine whether the system is in a steady state—was estimated at 0.86, indicating stationarity.

Second, we computed the kernel norms, which quantify how many new events in one stream are triggered by a single event in another stream. The results are visualised in Figure 14.

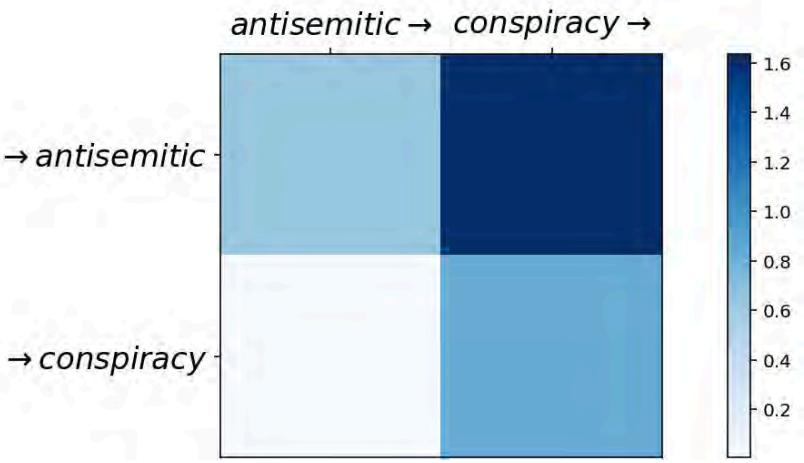


Figure 14. Estimated kernel norms between event streams. Kernel norms quantify the expected number of new events in one stream triggered by a single event in another stream. The heatmap displays these cross-excitation effects.

Our findings show that:

- **One conspiracy-related tweet triggers, on average, 1.6 anti-Jewish tweets.**
- One anti-Jewish tweet, in contrast, triggers only 0.33 additional anti-Jewish tweets.
- Anti-Jewish tweets have minimal causal influence on conspiracy content, with the triggering effect being weak and short-lived (typically fading after 1–2 hours).
- Conspiracy tweets have a sustained and powerful influence on anti-Jewish discourse, with a long-lasting triggering effect.

Both types of content demonstrate self-excitatory dynamics, but conspiracy content appears more self-sustaining, operating close to what is known as an explosive regime: a state in which the content could theoretically continue to escalate without external input. This suggests that even a small number of highly engaged users can generate and sustain a disproportionately large amount of anti-Jewish content when driven by conspiracy narratives.

4.1. The contagion of belief at user-level

To further investigate Aim 3, we adopted a user-centred approach to understand how engaging with the false claim that the Adass synagogue attack was staged or self-inflicted may shape online behaviour. This study is methodologically innovative in two important ways. First, we identified a group of users who actively promoted conspiracy narratives—either by retweeting existing conspiracy content or by posting their own messages after sharing such content. We compared their behaviour to a similar group of users who did not engage with conspiracy theories but did post anti-Jewish content. Second, we controlled for each user's behavioural patterns prior to exposure, which allowed us to distinguish changes driven by conspiracy engagement from users' existing tendencies.

We began with a dataset of X posts (including tweets and retweets) related to the Adass event, created by 14,337 Twitter/X users. From this group, we identified two distinct subpopulations:

- **Amplifiers (345 users): individuals who shared or posted conspiracy-related content.**
- **Non-Conspiratory users (13,609 users): individuals who did not engage with any conspiracy narratives.**

To ensure a fair comparison, we paired each amplifier with users from the non-conspiratory group who had similar behavioural profiles before the Adass event, and specifically, similar levels of anti-Jewish posting. Each amplifier was matched with 50 peers whose prior behaviour was most alike.

To build a manageable and representative sample, amplifiers were grouped into deciles based on how much anti-Jewish content they posted prior to exposure. From each decile, up to 10 amplifiers were randomly selected. Each was then matched with one user from the non-conspiratory group, resulting in a final sample of 58 amplifiers and 60 matched control users.

Recognising that people's online behaviour varies widely, we accounted for individual differences that could influence both their likelihood of engaging with conspiracies and their subsequent behaviour. We analysed each user's activity over the six months before the Adass event, including how active they were, the themes they engaged with, and the presence of anti-Jewish or toxic content in their posts. This helped us establish a behavioural baseline for each user, allowing us to more confidently isolate the impact of exposure to conspiracy content.

Our analysis focused on two key outcomes:

1. **Average toxicity per user – capturing the degree to which a user's language became more harmful or inflammatory.**
2. **Proportion of anti-Jewish messages per user – measuring the extent of anti-Jewish engagement.**

These measures allowed us to assess whether users who amplified conspiracy theories went on to produce more anti-Jewish or toxic content compared to similar users who did not.

The results show a clear and statistically significant effect of conspiracy engagement on user behaviour. On average:

- The proportion of anti-Jewish posts increased by 10.6 percentage points for amplifiers compared to matched controls.
- The average toxicity level of posts increased by 3.5 percentage points.

Both effects are statistically significant at the 5% level and remain robust even after accounting for individual users' prior behaviour. This suggests **that amplifying conspiracy content does not simply reflect users' existing biases but can actively contribute to more harmful and anti-Jewish discourse online.**

5. CONCLUSION

This report shows that online hostility targeting Jews, Zionists, and Israel increased sharply after 7 October 2023. This pattern is visible across multiple classification approaches, including toxicity and identity attack targeting Jews, Zionists, and Israel, as well as community-informed measures of old and new antisemitism. The size of the increase varies by category, with the largest increases observed in anti-Israel and anti-Zionist discourse and in new antisemitism. However, the core finding is consistent: whatever classification approach is used, **7 October 2023 marks a clear structural break in online hostile discourse in Australia.**

The Bondi terrorist attack on 14 December 2025 produced a further increase, although smaller and more localised than the post-7 October escalation. This increase was particularly visible in identity attack measures, suggesting that **the Bondi terrorist attack generated forms of group-directed hostility targeting Australian Jews.** Importantly, the Bondi attack occurred within an already elevated online environment, not a return to the pre-October 2023 baseline.

This report also shows that online and offline hate cannot be treated as separate domains. Real-world incidents, especially physical violence and symbolic aggression, can trigger measurable spikes in online anti-Jewish hate. Conspiracy narratives also play a major role. In the Adass case, false flag claims did not simply correlate with anti-Jewish hate, but helped sustain and amplify it. These findings suggest that **online hate can become reactive, self-reinforcing, and connected to offline harm.**

AUTHORS BIO

Matteo Vergani

Matteo is director of the Tackling Hate Lab and Associate Professor in Sociology at Deakin University and specialises in radicalisation and hate crime, publishing in leading international journals and securing large research grants. He collaborates with numerous institutions and government agencies in Australia and Canada. His research advances the systematic consolidation of knowledge in hate and extremism studies through large-scale systematic reviews and the development of rigorous measurement tools of online and online hate and radicalisation. His research programme fosters multidisciplinary collaboration across social sciences, data science, econometrics and engineering, leveraging advanced technologies for analysing digital archives and social media big data.

Andrea Giovannetti

Dr Andrea Giovannetti is Co-Director of the Tackling Hate Lab, Assistant Professor of Economics at the Australian Catholic University and a member of the Violence Research Centre at the Institute of Criminology of the University of Cambridge, where he previously held a Marie Curie Postdoctoral Fellowship. His research on organised crime, contemporary extremism and social cohesion combines machine-based econometrics with advanced computational methods in network theory to support policymakers and security agencies on a large spectrum of inter-connected issues. Andrea's collaborations with public stakeholders on complex social threats include London Metropolitan Police, Merseyside Police, Home Office and Home Affair.

Stephanie Zi Xin Ng

Stephanie holds a PhD in Engineering from Deakin University's Institute for Intelligent Systems Research and Innovation (IISRI). Her research focuses on advancing computational methods that leverage Natural Language Processing (NLP) and Large Language Models (LLMs) to address pressing social challenges. Her work spans a variety of topics, including the policy framing of contemporary social issues, stance detection in parliamentary debates, and agent-based modelling of inoculation strategies against online extremism. Her recent work explores video and image processing, including multimodal analysis of the privacy paradox in short videos and Computer Vision (CV) for hazard detection.

Haily Tran

Haily is a mixed-methods researcher in social psychology, focussing on psychological drivers of online radicalisation, violence, and hate-based ideologies. Her PhD examined the role of masculinity and victimhood in the mobilisation of Australian men toward far-right extremism. Experienced in experimental research designs and evidence-based practice, Haily has also contributed to multiple projects in hate crime prevention and countering violent extremism (P/CVE). She currently serves as HDR/ECR coordinator for the AVERT Research Network and is a member of the Australian Psychological Society's College of Forensic Psychologists.

Dan Goodhardt

Dan is a researcher at Deakin University, where he has collaborated for over a decade on studies of hate crime and violent extremism in Australia and Southeast Asia. He has extensive experience in collecting anti-Jewish hate data through the Jewish Community's Community Security Group and currently works as a senior manager with Victoria Police. As co-convenor of the Practitioners Working Group on Tackling Hate in Victoria at the Centre for Resilience and Inclusive Societies, he connects security practitioners with researchers to advance efforts in countering hate and extremism.

Huu Phuc Hong (Felix)

Felix is a data scientist and research assistant working at the intersection of computational social science and natural language processing. He was a highest-achieving graduate of Deakin's Master of Data Science (across both the Standard and Professional programs), and his work focuses on building reliable, transparent measurement tools for analysing online harms at scale. Felix develops hate speech detection approaches that combine large language models with smaller supervised classifiers, and creates end-to-end tooling to collect, organise, and monitor social media data. He also applies statistical validation methods and graph-based network analysis to map communities, narratives, and information flows, with an emphasis on clear, defensible evaluation.

Yinsong Cheng

Yinsong is an AI researcher and software engineer with a PhD in Computer Science and Electrical Engineering from Deakin University. His research focuses on trustworthy machine learning, large language model alignment, hate speech detection, explainable AI, uncertainty-aware learning, and scalable data analytics. He has developed LLM alignment methods for detecting implicit and context-dependent harmful language, with a particular focus on the Direct Preference Optimization method family. He also has experience across computer vision, time-series forecasting, probabilistic deep learning, and real-time data analytics. His work aims to build reliable, scalable, and interpretable AI systems for safety-critical and socially impactful applications.

Kewen Liao

Kewen is co-director of the Tackling Hate Lab and Associate Professor in Data Analytics and Machine Learning at Deakin University. Kewen is an accomplished algorithms researcher with a PhD in Computer Science. His expertise spans data science, machine learning, and theoretical computer science, with a focus on algorithm design and analysis, clustering and graph mining, time series and streaming analytics, and image and text data analysis. He is a Chief Investigator on an Australian Research Council Discovery Project and previously led a Defence Next Generation Technologies Fund Project. He also holds a US patent with Canon Inc. for a novel object-matching method in computer vision. Kewen is committed to cross-disciplinary research, leveraging data science and AI to drive significant societal and economic impact. He also holds key leadership and professional roles in leading national and international data science and AI conferences.

www.tacklinghate.org