Meta Oversight Board – CETaS Public Comment (Nov 2025)

Background

This submission response is from The Alan Turing Institute's Centre for Emerging Technology and Security (CETaS), prepared by **Sam Stockwell** (Research Associate at CETaS). The Alan Turing Institute is the UK's national institute for data science and artificial intelligence. CETaS' mission is to strengthen UK security through pioneering research on emerging technologies.

CETaS is conducting a project under the UK AI Security Institute's Systemic AI Safety Grants Programme exploring the role of AI-enabled information threats during security incidents and crisis scenarios. We will publish the report in early 2026, which includes analysis on the role of such threats during the Israel-Iran conflict in June 2025. We therefore welcome the opportunity to provide evidence to the Meta Oversight Board's case on this topic.

The role of AI-generated mis/disinformation during the Israel-Iran 2025 conflict

As the Israel-Iran conflict was unfolding in June 2025, we observed how generative AI tools were exploited by grassroots users and public officials on both sides for various purposes.

In the vast majority of cases, AI-generated content was designed to exaggerate the military strength of both countries and shape the public narrative over which side was deemed to be 'winning' the conflict as it unfolded. On the Iranian side, users were circulating synthetic images which appeared to show the destruction of Israeli F-35 fighter jets by Iranian forces, as well as similar images depicting dozens of missiles falling on Israeli cities. In one particular case, Iran's Supreme Leader (Ali Khamenei) posted an AI-generated image of missile strikes accompanied by a verse from the Quran, which received over 6 million views.

Similarly, Israeli officials shared content including an AI-generated video version of the bombing of Iran's Evin prison, which depicted what looked like a surgical strike blowing apart the entrance gates. However, in reality, the attack resulted in much more significant collateral damage and led to the deaths of civilians. Given that this clip was posted by Israel's Foreign Minister, there are concerns that the content was designed to undermine or detract attention away from credible claims of human rights violations after the strike was carried out. At the grassroots level, some users also shared synthetic videos of pro-Israel protests in Tehran, claiming that they showed mounting dissent against the Iranian regime and should inspire others in the country to carry out similar activities.

In analysing these various examples, we found that AI tools had become subsumed within the wider wartime propaganda machine for both countries during the conflict. The ability to use this technology to create realistic content that undermined perceptions around the enemy's military prowess, as well as emotive content which could evoke feelings of patriotism among the domestic population, assisted both Israel and Iran with morale-boosting efforts at home to continue the war. At the same time, it

also validated ambitions to carry out retaliatory action and contributed to wider psychological warfare operations conducted by non-AI techniques.

Outside of these cases, we also saw a handful of worrying examples where other nations were being implicated in the conflict as active collaborators through AI-generated videos. This included a deepfake of Pakistan's Defence Minister claiming that Islamabad had forewarned Tehran about the impending Israeli attack, as well as AI-generated clips falsely claiming that Nigeria was sending troops to Israel on a peacekeeping mission. If this type of content had gained more traction, it could have posed complications for preventing conflict escalation and a worsening of geopolitical tensions.

Finally, alongside the generation of content, we identified several cases where AI chatbots embedded on social media platforms helped to spread false information, sow confusion over the truth and enhance the credibility of synthetic media. For example, X's in-built chatbot Grok insisted that various AI-generated videos during the conflict were authentic, as well as revealing significant inconsistencies in its fact-checking capabilities when queried by users on certain posts. In one example, AI-operated X accounts of Perplexity and Grok seemed to suggest that claims of China providing active military support for Iran were valid. Finally, following AI-generated videos of missile strikes on Israeli cities, Grok also claimed that these were genuine. This created further uncertainty for panicked civilians trying to access information about safe zones during the conflict.

Prevalence and impact of AI-generated mis/disinformation during armed conflicts

Beyond the Israel-Iran conflict, CETaS has also been monitoring other types of recent security incidents – such as terrorist attacks, protests and violent riots – to understand what role AI-generated mis/disinformation has played during them.

Following the Southport murders in September 2024, several pieces of content circulated in the first 24 hours showed indicators of AI generation. Alongside AI-generated songs combining references to Southport with xenophobic content, similar cartoons promoted caricatures of Muslim men as a threat to British women and visuals of weapons designed to incite violence – including on Meta's own platforms. With recent analysis showing that xenophobic generative AI material containing violent narratives was three times more likely to be viewed than non-synthetic content during the Southport riots, there are concerns about large volumes of users being exposed to material that could incite offline harm targeted against certain demographic groups.

The brief India-Pakistan conflict in May 2025 saw similar methods being deployed on the international stage. Here, synthetic images of dead Pakistani bodies and militant figures were misleadingly presented as 'proof' of India's attack, with the intentional aim of justifying violent extremist discourses. Yet in addition to these more inflammatory pieces of content, others sought to undermine Pakistan's military response. For example, a deepfake of Pakistan's prime minister seemingly conceding defeat was circulated in parallel to AI-generated videos of alleged battlefield footage showing drones invading Pakistani territory.

In contrast to these crises, we observed how AI-generated disinformation was used for very different purposes during other security incidents. After the Charlie Kirk shooting in September 2025, blurry images released by the FBI showing the "person of interest" in their investigation were altered by users via AI upscaling tools to produce what were described as 'high-quality' versions of these original photos. However in many such cases, the AI-upscaling tools made incorrect inferences and added unverified details about the suspect, even changing key details such as the individual's facial structure and shirt. Such misleading details about the suspect could complicate law-enforcement operations and lead to the targeting of unrelated individuals who are misidentified by the public. Indeed, Grok even falsely claimed that a Utah-based registered Democrat had been identified as the shooter, wrongly attributing the information to major news outlets.

Challenges in detecting, labelling or fact-checking AI-generated content, and the effectiveness of policy, product and enforcement responses

Given the relatively novel nature of AI-generated mis/disinformation threats in these contexts, there is more research needed on the effectiveness of different responses. However, we have identified some approaches which may be promising for Meta to explore further. This includes solutions that embed fact-checking services directly into social media and messaging apps. In Taiwan, the messaging app LINE includes an in-app chatbot that allows users to submit website links for analysis and verification against previously fact-checked content. If the chatbot cannot match any of the existing data, users can forward their message for manual fact-checking. During the COVID-19 pandemic, the bot service handled more than 230,000 user submissions on potential health-related disinformation topics.

During the Indian elections held between April and June 2024, Meta itself launched a similar in-app service on WhatsApp known as 'Tipline' which allowed users to understand whether videos or audio notes they were viewing on the platform were AI-generated. Available in multiple Indian languages and English, the initiative used a combination of AI detection tools and human experts to ensure a robust and explainable process. As we have recommended in our previous research on AI and election security, further testing and integration of these services on apps where AI-generated content is prevalent could help to empower users to better determine the authenticity of content they encounter online.

We welcome Meta Oversight Board's attention to these important issues and are available for further comment if required.

Contact: Sam Stockwell, sstockwell@turing.ac.uk