Picture credit: Christopher Furlong/Getty Images/X

RIOT STARTER

Sam Stockwell *reveals how AI can add fuel to the fire in crisis events*

This article explores how AI information threats, such as deepfakes, can contribute to real-world harm during crisis events – including terrorist attacks, violent riots and international military confrontations. As AI tools become increasingly accessible in generating extremely realistic content at speed and scale, information threats pose even greater risks to public safety. This is particularly problematic during moments of crisis, where information voids could be exploited by threat actors to sow confusion and incite violence. With more users also resorting to AI chatbots for news consumption, AI tools are actively shaping public perceptions after crises which, in some cases, has involved the amplification of misinformation. Yet despite the many risks from AI within this context, there are also opportunities to use these tools for strengthening crisis response efforts in future incidents. Moving forward, it will be vital to continue monitoring this nascent threat landscape, while also exploring novel ways AI could enhance societal resilience.

In 2024 the UK witnessed a wave of violent riots and public disorder across the country following the murder of three young girls in the town of Southport. In analysing the factors which contributed to this violence, a parliamentary inquiry concluded that the absence of factual information on the identity of the suspect at the time "created a vacuum where misinformation was able to grow". In turn, these rumours were successfully exploited by extremist groups to mobilise and organise anti-Muslim protests – leading to several targeted attacks against migrant communities, multiple injuries to police officers and days of violent unrest.

Yet amid these unfolding developments, another type of information threat was receiving far less attention in the media coverage. While false information was being circulated on social media platforms, AI content generators were being weaponised to further exacerbate community tensions and incite violence. These included vitriolic AI-generated songs calling for UK citizens to "hunt down" groups associated with the perpetrator, as well as xenophobic AI-generated images of white men wearing Union Jack T-shirts chasing several males caricatured as Muslim stereotypes.

AI content generators can be weaponised to further exacerbate community tensions and incite violence

However, the events in Southport were just the beginning. Since then, we have witnessed at least 14 other crisis events where AI information threats have contributed to sowing confusion, spreading conspiracy theories and encouraging real-world harm.

Before delving into the specific ways that AI tools have contributed to information threats during recent crisis events, it is important to zoom out and identify the stages of the broader 'content life cycle' where these threats materialise.

AI tools are shaping the way content gets generated, disseminated and consumed by users on digital platforms. Firstly, AI content generators can be exploited by threat actors to rapidly create fabricated evidence from the scene of an incident or inflammatory content, which could incite violence. Secondly, AI-powered news aggregators – which scrape data from social media and repackage trending topics into polished news stories – can quickly be disseminated to large numbers of people through bot networks. Such content gives a strong impression of credibility and authority, which – when combined with rapid amplification – is able to actively shape public perceptions. Finally, AI chatbots often unintentionally scrape and reference viral social media posts (including AI-generated content) when queried on live incidents. This risks reinforcing the credibility of the original fabricated evidence in a highly personalised format.

Deepfakes are nothing new: from efforts to damage the reputation of political candidates during election campaigns to tricking employees into transferring money to fake CEOs, it is clear that synthetic content can cause real-world harm in a variety of sectors. However, we are now seeing threat actors explore this misuse of AI in other areas to further harmful agendas. After the Bondi Beach terrorist attack in Australia in late 2025, the pro-Kremlin Pravda disinformation network was found to have shared a deepfake image of one of the survivors who appeared to be 'staging' his injuries. Seen more than 10 million times on social media, the image perpetuated antisemitic conspiracy theories that the incident was a "false flag" attack orchestrated by Israeli intelligence agencies. Similarly, during the Israel-Iran conflict in 2025, AI-generated videos were circulated by both sides suggesting missile strikes were taking place in certain cities. Yet in doing so, this may have added even greater uncertainty for panicked civilians trying to access information about safe zones in the middle of aerial bombardments.

In an entirely different case, following the Charlie Kirk shooting in the US, blurry images of the suspect released by the FBI were fed into AI upscaling tools by so-called 'internet detectives' to produce what were described as "high-quality" versions. However, in many of these versions, the AI tools made incorrect inferences and added unverified details about the suspect – even changing key details such as the individual's facial structure and shirt. If left unchecked, these AI-edited versions could risk people being falsely identified, accused and potentially even targeted in future incidents.

Threat actors have long recognised that creating harmful content is only the first step towards achieving success in the information environment: once that material is ready, it ideally needs to be shared with as many users as possible. Here again, we see AI tools offering new opportunities during crisis scenarios.

AI TOOLS CAN CONVERSELY BE EMPLOYED TO STRENGTHEN CRISIS RESPONSE EFFORTS

Fake bot account networks have been uncovered in the context of recent elections taking place in the UK, US and Europe. Designed to mimic human behaviour and share material with unsuspecting users, these accounts can massively amplify the reach of disinformation campaigns. Yet alongside elections, we are also observing their presence in the immediate aftermath of major security incidents. Following a car ramming incident during Liverpool FC's Premier League victory parade in 2025, 40 percent of the speculation claiming that the suspect was an immigrant or asylum seeker came from inauthentic accounts that appeared to be interested in leveraging engagement to earn revenue.

After the Southport attacks in 2024, false information about the suspect was originally published by an AI-driven news aggregator website called 'Channel3Now'. Since being discovered, evidence has emerged that the site was set up via a service which markets itself as using AI to generate content for users seeking passive income. Perhaps most concerningly, there appeared to be minimal human editorial oversight over the content being spread by this site. In future crises, competition between these self-styled AI-augmented 'news sites' could drown out sources from legitimate outlets. In turn, we may see a 'race-to-the-bottom' where sensationalist headlines are prioritised over factual accuracy in order to generate revenue, even if they cause confusion or promote divisive narratives.

Beyond AI content generators and aggregators, chatbot interfaces are now often integrated by default into our social media platforms, internet search engines and private messaging apps. This accessibility and proximity is resulting in users increasingly relying on these services for news consumption, including during live incidents when information is often scarce or uncertain. Typically, chatbots are able to search through their training data for information on historical events. However, if information on an unfolding crisis was not included during the training stage or does not yet exist, it must rely on the internet for support. Inevitably, this leads to situations where models will scrape poor quality content, misinformation or deceptive advertising to fill in knowledge gaps. Yet instead of acknowledging their limitations or declining to weigh in on sensitive topics, models will happily oblige with user requests

and repackage inaccurate information in a highly personalised and authoritative-sounding way.

During the India-Pakistan conflict in 2025, X's in-built Grok chatbot wrongly identified old video footage from a Sudanese airport as an Indian missile strike on a Pakistani airbase. In the middle of a live conflict, this does no favours in mitigating the risk of further military escalation. After the Charlie Kirk shooting, Grok went so far as to incorrectly identify an unsuspecting individual as the confirmed perpetrator, leaving them "shocked" by the episode. As with the AI upscaling tools referenced earlier, there are serious concerns moving forward that these types of misattributions by AI chatbots could lead to individuals with no connection to an incident being wrongfully targeted or detained.

IT IS ABUNDANTLY CLEAR THAT SYNTHETIC CONTENT CAN CAUSE HARM IN A VARIETY OF SECTORS

Despite the many different ways that AI can be weaponised by threat actors during crisis events, it is also important to note that these same tools can strengthen crisis response efforts. Novel data analytics platforms such as the Social Media Analytics and Reporting Tool (SMART) are incorporating AI classification techniques to help with improving situational awareness and sentiment analysis during live incidents. Following the Southport riots in 2024, SMART was able to uncover how quickly false information relating to the incident was spread on social media, as well as UK hotspots where these misleading narratives were geographically concentrated.

Additionally, while AI chatbots have shown various issues with fact-checking content in the context of live events, researchers are also training bespoke models which could help to counter disinformation at similar speed and scale. 'DebunkBot' is an example of such a model. By processing vast quantities of knowledge on different conspiracy theories, DebunkBot aims to counter them with persuasive, tailored rebuttals. When tested on 2,000 participants who expressed a belief in at least one conspiracy theory analysed, conversations with the chatbot reduced their confidence in such theories by 20 percent, with some reductions lasting up to two months.

There is great deal of uncertainty when considering whether AI will help or hinder us during future crisis scenarios. On the one hand, we have seen several concerning cases where these tools have allowed threat actors to rapidly distort public perceptions, spread harmful conspiracy theories and even promote real-world harm. Yet on the other, AI-powered tools such as SMART and DebunkBot show that this same technology holds promise in tackling long-standing challenges at the intersection of disinformation and violence. Only by having frank conversations about how we can unlock these advantages, while introducing necessary safeguards, will we ensure that AI can play a role in enriching our democracy – rather than undermining it ●

Sam Stockwell
is Senior Research Associate at the Alan Turing Institute's Centre for Emerging Technology and Security.

Chatbot interfaces can cause chaos during live incidents when information is often scarce or uncertain

Credit: Crown Copyright