Irregular Warfare
PRISM Perspectives

The Blind Spot in Defense: Inside Counterterrorism's Black Box

By Shaheer Ahmad

17 September 2026

Shaheer Ahmad is a Research Assistant at the Centre for Aerospace & Security Studies, Islamabad, Pakistan. His research interests include emerging technologies, geopolitics, and military strategy. He can be reached at ahmed.shaheer@casstt.com.

On July 10 2026, a Cambridge University researcher published a report that mapped the use of frontier artificial intelligence (AI) models, including ChatGPT, Claude, Grok, DeepSeek, Gemini and Meta AI, by terrorist groups in Western Africa. Based on the post-hoc testimonies by the demobilized members of the Boko Haram faction, the terrorist organization [with the assistance of Islamic State in Iraq and Syria (ISIS)] employed iterative prompting to obtain incremental assistance on innovating improvised explosive devices (IEDs), improving surveillance, building weaponized drones and assisting with troubleshooting. While Cambridge University analysts have pointed toward the risks these models pose in the hands of terrorist groups, an under reported factor merits attention: the black box problem, which has less direct visibility into how a model thinks and process the iterative prompting.

The AI black box problem refers to the inability of humans to see how a deep neural network makes decisions or arrives at a particular conclusion. Generally, language learning models (LLMs) process numerous variables simultaneously to find geometric patterns in multi-dimensional spaces. In this high-dimensional environment, the algorithms draw inferences from an infinite number of permutations that could deviate the model from the earlier inputs that informed its decision-making. Similarly, modern LLMs are trained across a wider gradient with billions of parameters inserted to inform their reasoning. This reasoning is computationally difficult for humans to ascertain, as it is dispersed across a wide spectrum.

Secondly, the protective guardrails of the AI models comprise four interconnected components--checker, corrector, rail, and guard. These four components ensure that generated content does not contain biased, harmful, or misleading information. Despite this regulatory promise by the AI developers, the opaqueness of AI models' internal workings resists straightforward human comprehension. Henceforth, these regulatory protocols alone do not remove the vulnerability of modern AI models. This blind spot amplifies a range of safety, security, operational, and accountability risks that are computationally overwhelming to track and resolve.

According to the US Department of Energy's (DOE) AI Risk Assessment 2024, AI systems are susceptible to data poisoning and evasion techniques, where adversaries have the potential to facilitate sabotage and conduct attacks against critical infrastructure. While the report makes a case against the AI's integration in critical infrastructure, it is equally applicable to the frontier models whose compliance and ethical protocols could be bypassed by non-state entities.

This underlying vulnerability was exploited by the Boko Haram operatives beginning as early as 2023. Using jail-breaking techniques such as prompt engineering, the specialist operatives successfully overrode the safety protocols of frontier AI models. According to the testimonies recorded by the Cambridge University report, the AI systems offered precise suggestions on repurposing weaponized drones by advising on payload

reduction and explosive release mechanisms. These responses were then cross-referenced by the commanders against multiple AI systems to omit discrepancies.

This prompt-based engineering approach directly probes the AI black box problem. The deep neural networks of AI comprise thousands of artificial neurons developed to solve a problem. This complexity creates a transparency problem, where the repeated trial-and-error techniques overwrite the system instructions and manipulate the pre-approved protocols. This active manipulation seemingly renders the safeguards and regulatory guardrails posing no significant hurdles to malicious LLM use. This further reinforces the threat that the AI outputs are hinged on the stated intent, revealing a structural vulnerability that could be exploited and manipulated by terrorist entities. According to a self-funded experiment by *Tech against Counterterrorism AI Benchmark*, *twenty seven* leading AI models were tested across 152 ways to assess the misuse of AI in terrorist activities. Around a third of responses provided real assistance toward making a real weapon or planning a mass attack. This demonstrates the risk that terrorist organizations could redesign the operational processes of AI models with more agility.

This kind of AI-assisted force optimization represents a shift in mission planning and execution strategies of terrorist entities. Contemporary frontier AI models are capable of providing inference and technical assistance that could uplift the capabilities of terrorist groups. This threat is further exacerbated by the presence of *open-weight* AI models. Closed-weight AI models [all frontier AI models] cannot be modified by end users. The *open-weight* models allow the user to customize, modify, and tailor the model according to their requirements. They are easily downloadable via creator repositories and online platforms. The Boko Haram members were trained by specialist operatives from ISIS on close-weight AI models. There is a higher probability that these operatives will be doing the same with other regional branches across the world. Besides this, the employment of open-weight AI models by terrorist groups couldn't be a distant option.

This creates a possibility of a global chain reaction across volatile regions, including the Middle East, South and Central Asia, aggravating already fragile security environments. This democratization of AI raises concerns for the governments and militaries involved in counter-terrorism efforts. Contemporary trends and empirical evidence suggest that such groups have demonstrated technological adaptation with a higher probability of capability fusion with other groups across the world. With increasingly dispersed command and control (C2), these groups have seamlessly utilized AI and other tools for tactical, operational, and informational purposes.

The availability of frontier AI models provides a cost-effective, agile assistant that could cater to operational demands of terrorist groups in the wake of political and resource-constrained environments. Attackers now design and test new prompting techniques before defenders can rewrite the safety protocols. Although AI systems are designed mainly to refuse certain queries from users. The Cambridge report showcases that jail-breaking and abliteration methods can strip or overwrite the safety protocols with minor inconveniences.

This threat is further exacerbated by the introduction of more intelligent and autonomous frontier AI models such as GPT-6 Astra. According to empirical evaluations, the model exhibited superhuman competency by outperforming previous frontier AI variants. The new model is more opaque than its predecessors, which raises significant security and ethical concerns. Thus, the precedence of computing efficiency over transparency worsens the black box paradox for policymakers.

The Boko Haram case serves as a wake-up call for contemporary security thinkers about the perils posed by AI in the asymmetric domain. The threat AI poses in the hands of terrorist groups demands transparent, auditable verification protocols along with systematic benchmarking to strengthen the existing guardrails. Moreover, constant red-teaming is crucial to promptly detect jailbreaking and abliteration techniques. As the AI matures, the decrease in transparency necessitates the policymakers to rethink black box problem with a cautionary tale.