Iran and the AI Threats We Cannot See

by Hamdi Malik **(/experts/hamdi-malik)**

Nov 4, 2026

ABOUT THE AUTHORS



Hamdi Malik (/experts/hamdi-malik)

Hamdi Malik is an associate fellow with The Washington Institute and cofounder of its Militia Spotlight platform.

Brief Analysis

A growing body of evidence shows how Iran and its proxies are using artificial intelligence and especially more easily manipulatable open-weight models as tools of warfare, repression, and propaganda.

AI companies' reviews of malicious use attempts of their products—such as Anthropic's September threat report **(https://www.anthropic.com/threat-intelligence-report-september-2026)** —are highlighting growing Iranian efforts to break through safety features in order to misuse AI models. Anthropic's report describing Iranian actors using Claude, Anthropic's family of AI models, for domestic surveillance, malware development, and research into targeting U.S. naval forces. Likewise, the report describes a weapons engineering cell in Houthi-held territory using Claude to develop guidance software. Its rocket test apparently failed, and Anthropic found no evidence of an operational weapon. But the cell used AI to investigate what went wrong and built an offline simulation toolkit that could survive the loss of Claude access.

While Anthropic detected and investigated these activities **(https://www.anthropic.com/threat-intelligence-report-september-2026)** because the actors used its closed-weight model subject to the company's safety guardrails and misuse monitoring, allowing Anthropic to terminate access, these reports also provide a window into how Iran may be deploying open-weight AI models, designed with greater end-user flexibility and less oversight.

Iran's interest in AI sits alongside an ambition for its use in offensive operations. In August, Mohammad Reza Naqdi, a senior adviser to the IRGC commander, said the force had been tasked with preparing to conduct operations on enemy territory far from Iran when ordered **(https://www.mehrnews.com/news/6914755/%D8%B3%D8%B1%D8%AF%D8%A7%D8%B1-%D9%86%D9%82%D8%AF%DB%8C-%D9%87%D9%85%D9%87-%D8%AA%D8%B3%D9%84%DB%8C%D8%AD%D8%A7%D8%AA-%D9%86%D8%B8%D8%A7%D9%85%DB%8C-%D9%85%D8%A7-%D8%A7%D9%85%D8%B1%D9%88%D8%B2-**

%D8%A8%D9%88%D9%85%DB%8C-%D9%87%D8%B3%D8%AA%D9%86%D8%AF) . Thus, these reports highlight a developing blind spot in the capabilities of AI lacking the stopgaps and monitoring deployed by a company like Anthropic, and how these models can shape the offensive capabilities of Iran and its proxies. Specifically, the increasing availability of powerful open-weight models that Iran can deploy on its own systems, outside a provider's monitoring and enforcement signals a potential new stage of its AI capabilities. Understanding the risks these models represent and how Iran may operationalize these models is key to highlighting the need for future stopgaps that can be deployed to slow Iranian advances in weaponizing AI technology.

AI on Iran's Own Terms

Open-weight AI models were developed to provide end users with greater customizability and control. With open-weight models, the AI company that created the model makes the model's weights—the numerical values that shape how the model learns and responds—available for download. This is in stark contrast to models such as Claude and ChatGPT, where weights are considered proprietary and are closely guarded trade secrets. Once the weights are downloaded, the entire model can run locally without an internet connection on any device that possesses sufficient computing power to handle the model.

With these offline models, the user's interactions cannot be monitored by the AI company and access cannot be remotely revoked—standard safety features regularly deployed by AI companies to prevent misuse of closed-weight models. And while AI companies build safeguards into their models designed to restrict harmful assistance even offline, it has been demonstrated that these protections can be successfully stripped from powerful, publicly available open-weight models.

Alvin Wang Graylin, a global technology strategist who advised the U.S. side ahead of U.S.–China AI talks (https://www.youtube.com/watch?v=Slle5_AxBzs&t=4167) in September, has argued that small AI models built or adapted for specific tasks—including open-weight models—can pose greater threats than larger, general-purpose models (https://www.thecipherbrief.com/the-biggest-ai-models-are-not-the-biggest-threats) due to these and other factors. And as a case in point, versions of Gemma 4, Google's family of open-weight models, appeared publicly with safety restrictions stripped out within 48 hours of the April 2026 release. (https://pbxscience.com/googles-gemma-4-gets-abliterated-just-days-after-launch/) Leading open-weight models are already approaching the leading closed frontier AI models on some intelligence tasks (https://www.anthropic.com/research/intelligence-targeting-conventional-weapons-capabilities) . Anthropic expects open-weight models eventually to reach capabilities demonstrated by today's frontier models in weapons-development tasks.

The world of open-weight models is therefore opening the door for Iranian teams to develop offline AI models altered to answer requests their original versions would refuse. Open-weight models provide an additional advantage via the process of fine-tuning. The IRGC and other Iranian military and intelligence institutions can fine-tune open-weight models locally on the specialized data and training material accumulated over decades, allowing these models to become better at solving problems specific to those institutions' operations.

Iran's () loss of nuclear scientists (https://apnews.com/article/israel-iran-nuclear-science-attacks-e298f00ba261debba4499a48c9df8b3d) and missile specialists (https://www.iranintl.com/202603236787) makes access to AI assistance particularly significant. Specialized open-weight models can allow Iranian research teams to apply existing expertise to new weapons projects, analyze test results, and work through design problems. This assistance could help close some of the gaps left by those losses of human expertise, although it would not replace the scientists' practical experience.[CC1] (//420F6A5A-578E-4836-A15F-55D7947EC4D3#_msocom_1)

Additional use cases already documented by Anthropic include attempts to develop and deploy malware

(https://www.anthropic.com/threat-intelligence-report-september-2026) for surveillance and propaganda purposes. In cyber operations, Iranian actors have already **used Claude for malware (https://www.anthropic.com/threat-intelligence-report-september-2026)** development. In July, U.S. agencies **warned of Iranian-affiliated attacks disrupting industrial control systems (https://content.govdelivery.com/accounts/USDHSCISA/bulletins/4219689)** across water, energy, and government facilities. Biological weapons research presents a further potential danger: **SecureBio's assessment of AI assistance with laboratory problems (https://securebio.org/benchmarks/)** suggests that as open-weight models gain these capabilities, Iran and its proxies will gain access to a tool that helps users understand and apply biological research. This new access can help troubleshoot laboratory experiments relevant to biological weapons development. Carrying out that work would still require laboratories, materials, and practical expertise, but even as proprietary models have become famously cautious around queries related to biological research, open-weight AI lacking equivalent safeguards can play a key role in accelerating and troubleshooting this research.

Developing a Portable Knowledge Network

Through **distillation, selected capabilities can then be transferred to smaller models by training the smaller models on outputs from larger, specialized model (https://research.google/blog/distilling-step-by-step-outperforming-larger-language-models-with-less-training-data-and-smaller-model-sizes/)** s, further economizing a limited amount of available GPU. These smaller models have the additional benefit of portability; files could be designed to be small enough to be copied onto USB drives and distributed within Iran's military and intelligence institutions, across its regional networks and to operatives around the world.

Terrorist groups face a dilemma: sharing practical expertise often requires contact between members, but those contacts can expose them to intelligence agencies (https://www.rand.org/pubs/monographs/MG331.html) . Restricting those contacts protects the organization but also hinders its ability to pass on skills. Specialized open-weight models that run offline on a laptop could reduce some of that dependence on personal contact. Through USB access, recipients can run smaller models on their own computers without ever establishing an internet connection that would make these models vulnerable to detection. These models could be designed to provide specialized knowledge and interactive reasoning tools that analyze problems, suggest solutions, and help revise failed approaches—functions that would otherwise require repeated, potentially monitorable access to an expert.

Access to unmonitored knowledge has major implications; these models could be designed to address questions about obtaining the materials and equipment needed to construct explosive devices and execute attacks without attracting suspicion or alerting the authorities, along with interpreting test results, investigating failures, and conducting reconnaissance. Their responses could be tailored to the circumstances of a planned attack, helping operatives assess options and revise plans as circumstances change—additional usability that manuals traditionally used to instruct these efforts cannot offer.

AI is unlikely to replace a handler's role in directing attacks, with this person-to-person communication remaining a vulnerability. Likewise, police could still discover models and saved conversations on a seized computer. But AI assistance can allow for fewer calls and messages seeking advice, making it harder for investigators to identify an operative's contacts and uncover a wider plot.

Addressing Chip Limitations

Iran will need sufficient computing power to achieve these results; locally hosted AI models require specialized chips, commonly called GPUs. Larger models—generally considered more powerful—demand more powerful equipment. U.S. sanctions are in place to prevent chips from leading supplier Nvidia from ending up in Tehran's hands. But smuggling can become a source of Nvidia chips, while **Chinese company Huawei's AI chips**

(https://www.huawei.com/en/news/2025/9/hc-xu-keynote-speech) have no such barriers in place to prevent Iranian purchasing.

In fact, Tehran already claims some access to such chips. The **() government-funded Simorgh computing center (https://www.cambridge.org/core/journals/iranian-studies/article/artificial-intelligence-in-iran-national-narratives-and-material-realities/9D2C33F3F1BC459712651443F1D203FB)** at Amirkabir University of Technology has advertised access to Nvidia H200 GPUs **(https://t.me/hpcrc_aut_news/912)** . The university has research agreements with Iranian defense organizations **(https://www.iranwatch.org/iranian-entities/amirkabir-university-technology)** and research links to the IRGC's Imam Hussein University **(https://www.iranwatch.org/iranian-entities/imam-hussein-university-revolutionary-guards)** . It also appears on the Pentagon's list restricting funded research collaborations **(https://www.cto.mil/wp-content/uploads/2026/07/FY25-Section-1286-List.pdf)** .[CC2] **(//420F6A5A-578E-4836-A15F-55D7947EC4D3#_msocom_2)**

Meanwhile, rapid competition to provide increasingly efficient AI models means the open-weight models Iran can run with its current hardware capabilities are becoming more powerful. These developments highlight that preventing access to chips is necessary but insufficient—either countries where open-weights models are being developed will put regulatory stopgaps on their availability, or these models will become increasingly accessible.

Conclusion

Iran's documented use of AI shows the danger is no longer theoretical, yet what has emerged so far may be only an early indication of the risks ahead. Powerful open-weight models are already freely available, and once released, they cannot realistically be recalled or brought back under their developers' control. But AI capabilities are also advancing rapidly, and future generations of models are likely to offer far more dangerous capabilities to hostile states and non-state actors.

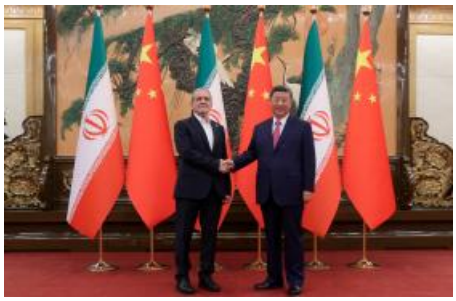
The fact that some powerful models are already beyond control makes it all the more important to prevent even more capable systems from becoming similarly accessible where possible while soberly understanding the risks the open-weights models that are released represent. Experts are already putting forward proposals to address these risks. For example, AI entrepreneur and Link Ventures founder David Blundin **has suggested tracking AI computing hardware and monitoring its use, drawing comparisons with controls on nuclear materials (https://www.youtube.com/watch?v=AxzcWOxzkiw&t=2629)** . This is one of several possible approaches that should urgently be examined, alongside other technical, regulatory, and international measures to determine what could meaningfully reduce the threat.

As both major developers of advanced AI models and producers of the computing hardware needed to run them, the United States and China will both have an outsized role in determining the accessibility of open-weight models. Bilateral AI talks in September established technical talks to take place in November, and the regulatory agreements or lack thereof that emerge from these talks will shape how easily Iran can access more powerful AI capabilities in the future—from both a model and hardware perspective. Current Iranian efforts on closed-weight models made visible in reports from companies such as Anthropic provide a window into the stakes of these discussions. Published attempts to misuse closed-weight models should be understood as the tip of the iceberg regarding Iran and its proxies' vision for the next stage of their AI-assisted offensive capabilities as newer, more powerful open-weight models loom on the horizon.

[CC1] (//420F6A5A-578E-4836-A15F-55D7947EC4D3#_msoanchor_1) I suppose their development can also mitigate the efficacy of targeting human specialists as models become more capable, not sure if you would want to add a sentence to that effect

[CC2] (//420F6A5A-578E-4836-A15F-55D7947EC4D3#_msoanchor_2) Patrick suggested adding details about the capabilities of Iran's technical universities to highlight their adeptness in incorporating technological developments.

RECOMMENDED



ARTICLES & TESTIMONY

China Won't Bail Out Trump's Lost War

Nov 3, 2026

◆

Henry Tugendhat

(/policy-analysis/china-wont-bail-out-trumps-lost-war)



BRIEF ANALYSIS

After MINURSO: Managing Tension, Preventing Conflict

Nov 4, 2026

◆

Sabina Henneberg

(/policy-analysis/after-minurso-managing-tension-preventing-conflict)



BRIEF ANALYSIS

Lebanon's Framework Is Stuck in Neutral: Political Will, Not Capability or Verification, Is the Key

Nov 3, 2026

◆

Assaf Orion

(/policy-analysis/lebanons-framework-stuck-neutral-political-will-not-capability-or-verification-key)