AMNESTY INTERNATIONAL

AMR 04/1128/2026

GLOBAL: RESPONSE FROM OPEN AI TO AMNESTY INTERNATIONAL

As part of its research into the human rights policies and practices of key technology companies in relation to Generative AI use in military contexts, on 5 May 2026, Amnesty International wrote to Open AI and Anthropic requesting information and a response on the human rights risks pertaining to military use of these technologies, as it has not previously corresponded with these companies on this topic. Anthropic has not responded at the time of publication of this document. Open AI sent a response on 12 May 2026, which is attached below.

OpenAI

1455 3rd Street
San Francisco, CA 94158

May 12, 2026

To:
[redacted]
Amnesty International

Thank you for your letter regarding OpenAI's human rights policies and practices in relation to generative AI products in military contexts. We appreciate Amnesty International's engagement on these issues.

In our April 15, 2026, letter, we expressed our view that democratic security and the protection of human rights are deeply interconnected, and democratic governments must be able to understand both the capabilities and risks associated with frontier AI systems. We believe that, with appropriate safeguards and meaningful human oversight, AI can help democratic societies protect people, deter aggression, and potentially reduce the likelihood of future conflict.

AI technologies are rapidly proliferating and are being developed and deployed by a wide range of state and non-state actors. The relevant question, then, is not whether AI will be used in conflict, but how to mitigate harm and uphold legal and human rights standards in an environment where such use is increasingly unavoidable. Though we believe in the potential benefits of AI for enhancing human security, the realization of those potential benefits will require rigorous safeguards, accountability, and ongoing scrutiny regarding how such systems are developed and deployed. **We therefore believe that engagement between frontier AI developers and democratic governments should occur within clear governance frameworks designed to uphold human rights, preserve accountability, and maintain meaningful human oversight.**

This belief is consistent with OpenAI's long-standing approach, which includes both policy-level restrictions on how our services may be used by commercial users, and model-level safeguards designed to reduce the risk that our models facilitate serious harm. These policies have long prohibited the use of our services for the automation of high-stakes decisions in sensitive domains like national security without appropriate human review. Additionally, OpenAI prohibits the use of its services to facilitate illegal activity, which includes those with severe harms such as war crimes, crimes against humanity, genocide, torture, or mass domestic surveillance. These restrictions apply to users and deployments of OpenAI services and are reflected in our Usage Policies and related governance materials.

OpenAI's commercial model behavior standards are designed to *prevent* our models from facilitating critical harms, including violence, terrorism, mass domestic surveillance, persecution, or other serious abuses. At the model level, OpenAI incorporates principles such as human safety, human rights, and human control into model behavior. As reflected in the Model Spec, these principles are intended to shape how models respond, including by preventing assistance that would enable critical harms and by reinforcing the importance of human oversight in sensitive or consequential contexts.

With respect to OpenAI's work with the U.S. government, OpenAI has publicly described three red lines that guide its work with the Department of War (DOW): no use of OpenAI technology for mass domestic surveillance, no use of OpenAI technology to direct autonomous weapons systems, and no use of OpenAI technology for high-stakes automated decisions. These redlines are included in our agreement with the DOW in addition to the prohibition against using OpenAI technology for illegal activity. Importantly, we do not rely on these contract restrictions in isolation. The agreement incorporates multiple, mutually reinforcing protections, including deployment limitations, safety controls, and oversight mechanisms. We will incorporate technical guardrails designed to refuse, limit, or redirect requests that would facilitate prohibited activity. Embedded staff will provide oversight and better enable

[redacted]

OpenAI

1455 3rd Street
San Francisco, CA 94158

identification and mitigation of emerging risks. Where we identify noncompliant activity, we will work with the Department to remediate. Together, these measures create practical constraints on misuse that go beyond contractual commitments.

At a broader level, OpenAI recognizes that questions surrounding AI and military applications are complex and evolving. We do not view these challenges as solely technical questions, but as issues that require continued engagement among governments, civil society organizations, researchers, and industry. We therefore appreciate Amnesty International's efforts to contribute to this discussion and welcome continued dialogue on these important topics.

[redacted]