On the escape of the OpenAI agents: “Reducing this to PR misses what actually happened”

Andreas Hepp on the OpenAI agents that escaped their test environment and attacked Hugging Face.

Interview conducted by Helmut van Rinsum for Horizont (horizont.net), July 27th, 2026. Translated from the German. The case in question: https://openai.com/index/hugging-face-model-evaluation-security-incident/

Late last week it emerged that two OpenAI models had escaped from a test environment and penetrated the infrastructure of Hugging Face. The incident coincides with the planned stock market listings of OpenAI and Anthropic, which has already led some media outlets to speak of a “PR stunt.” How do you assess the incident?

The claim that this is a “PR stunt” misses the multilayered character of the current situation. Long before OpenAI existed, there was a debate about AI safety in AI contexts, and the scenario of an AI agent “escaping” was already discussed there as one of the risks. OpenAI was founded with people from these contexts, extending to effective altruism and rationalism, where an escaping AI is regarded as the greatest existential risk to humanity. Many of these people have since left OpenAI and now work at Anthropic or at other non-profits, and part of this is OpenAI’s development into a commercial company. In doing so, Sam Altman has repeatedly drawn on the debate about an artificial general intelligence (AGI) or “super intelligence,” and on the fears associated with it. Fear, too, can be a motivation to invest, or to be the first to command a technology. But reducing this to PR misses what actually happened, both technically and socially. We are dealing with something more complex here. The sequence of events also speaks against a calculated move: Hugging Face went public on July 16 without knowing who was responsible, and had alerted law enforcement. OpenAI found the traces in its own systems only over the following weekend. Had this been staged, Hugging Face would have had to play along.

OpenAI itself speaks of an “unprecedented cyber incident” and has emphasized the capabilities of its own model. How is this choice of words to be interpreted?

At OpenAI, the term “unprecedented” is meant in the first instance in a cyber-technical sense – the wording is “unprecedented cyber incident with advanced cyber capabilities.” But it also refers to the debate I have just described. What is meant is that the escape of agents has so far been discussed in theory, whereas we are now experiencing that it can actually happen. It is not entirely unprecedented, however: a benchmark paper from May already regarded the autonomous development of exploits as real, and one day before this post OpenAI itself had reported a second escape from a test environment. In the language of the AI safety community, the models search for a “path of their own” to complete the original task – this is referred to as “specification gaming” or “reward hacking” – and this can involve something entirely different from what was intended on the human side. That is precisely what happened here: the models were fixated on solving the benchmark, and they obtained the solutions from Hugging Face. This was not about self-preservation but about a detour toward the goal – and the test environment did not hold. With its post, OpenAI connects to this discourse, and yes, this also stands for the capability of its own model. But it stands as well for the fact that a scenario so far discussed only in theory can actually occur. Part of the picture, however, is that the safety filters had been deliberately switched off for this test: OpenAI runs such evaluations without the classifiers that prevent risky cyber activities in production.

Do you see a recurring pattern here in the way tech companies turn narratives of losing control into signals of competence?

What I have said so far shows that we are not dealing simply with a narrative. There is narrative work nonetheless: OpenAI writes that its own security team detected the anomalous activity internally. According to recent reporting by Reuters, however, OpenAI noticed nothing for about a week and made the connection only after the post by Hugging Face. Even if OpenAI disputes parts of the Reuters account, a failure of its own monitoring is thereby turned into a demonstration of capability. OpenAI does communicate strategically, then, which conversely does not mean that we can reduce the incident to this. The underlying problem remains untouched by it. In the AI safety community it is discussed under the heading of "alignment": the question of how such systems can be aligned with human values.

Can outsiders tell at all whether an incident of this kind is a genuine security breach or a strategically placed message?

Companies like OpenAI are not closed worlds. In the San Francisco Bay Area, developers are in closer contact than one might think. How the leadership presents itself in public has been contested again and again, internally as much as externally. In this case, moreover, another central developer of AI was affected in Hugging Face. Even if one certainly does not know everything, the scale of such an incident can be assessed: the sequence of events has by now been reconstructed via Reuters from people close to the investigation, and it diverges from the company's own account. Added to this are the disclosure by Hugging Face itself, public statements by its co-founder about the period of the attack, and the technical reconstruction carried out within the scene itself. Conversely, the case also shows where the limits lie: the accounts given by Hugging Face and OpenAI do not yet fully agree on the initial point of entry, and both investigations are still ongoing.

In your view, how should a tech company act when an incident like this occurs – is a blog post the most sensible approach in PR terms?

OpenAI did seek contact with Hugging Face – but late. The attack ran from July 11 to 13, Hugging Face went public on July 16 and had alerted law enforcement, and OpenAI got in touch on July 20. That Hugging Face detected the attack independently, as OpenAI itself writes, is correct in this sense: there was no alternative to it. That the labs and companies developing frontier models communicate externally through posts and blog entries is their usual way. The world learned about ChatGPT through a blog post as well, and through the network then still called Twitter, and OpenAI was itself surprised by the reactions. So communicating with those affected and with the public is certainly a good starting point. The real discussion at present, however, is a different one, and this interview stands for it too: What part did the blog post by OpenAI play in the overall course of the incident? To what extent has the incident brought about what has always been warned of with regard to AI? And what consequences must we draw from it for the development and regulation of AI, especially here in Europe? In the light of such questions, two points seem to me particularly relevant. The first is a regulatory gap: the incident occurred during the internal pre-deployment testing of an unreleased model. The obligations of the AI Act, however, are tied to "placing on the market," and how a company secures and monitors its own test environments is effectively unregulated. The second is an asymmetry: Hugging Face was unable to use the commercial frontier models for the forensic analysis, because their safety filters rejected the attack data, and fell back on an open Chinese model running on its own infrastructure. Anyone who wants to protect themselves therefore needs capabilities they can operate themselves – and for Europe this is not a question of industrial policy but one of security.

Prof. Dr. Andreas Hepp is Head of ZeMKI, the Centre for Media, Communication and Information Research at the University of Bremen (https://zemki.uni-bremen.de), and co-directs the DFG/FWF Research Unit 5656 "Communicative AI" (https://comai.space). He conducts research on Silicon Valley; his latest book on the subject, "Pioneer Communities: How Our Digital Futures Emerge in

Silicon Valley and Beyond," will be published by Polity Press in December 2026 (https://www.politybooks.com/bookdetail?book_slug=pioneer-communities-how-our-digital-futures-emerge-in-silicon-valley-and-beyond--9781509559619).