ISSUE 8 · AI GUARDRAILS AS A BLOCKLIST

Guarding the guardrails: AI reached for a 2005 idea

KEY TAKEAWAYS

- On July 1, 2026 Anthropic redeployed Fable 5 globally, with Mythos 5 restored to its Project Glasswing partners, after Commerce lifted the export ban; Fable had launched June 9 and was suspended days later when the controls hit, following a reported bypass that let it generate exploit code.
- Amazon showed Fable's safeguards could be bypassed with a particular prompt framing, and Anthropic's fix was a new safety classifier that targets and blocks that framing, catching the technique in more than 99% of cases; NIST's Center for AI Standards and Innovation tested it and called the safeguards 'extraordinarily strong.'
- CAISI signing off on a specific fix as the condition for restoring access is new; a government body approving one patch before an AI model could ship again has not happened before.
- The fix is a filter that reads the incoming prompt and decides whether the words look like an attack, which is a learned, probabilistic blocklist, the oldest mitigation in application security and the one with the longest track record of failure.
- History repeats: SQL injection denylists always got bypassed (comment insertion, case variation, hex and Unicode and double encoding, concatenation) and were solved structurally by parameterized queries that separate code from data; WAFs are signature engines walked around for twenty years, with a sensitivity tradeoff inherent to matching against a list of bad patterns.
- AI gets no structural fix because a language model has exactly one channel: instructions and data arrive as the same stream of natural language with no architectural place to put a boundary, so the instruction and the injection are made of the same material.

> **"The instruction and the injection are made of the same material."**

We have run this experiment before

Blocking bad input by recognizing bad input is the oldest and most-failed mitigation in application security. SQL injection denylists were endlessly bypassed and only solved structurally by parameterized queries; WAFs have been walked around for twenty years and carry an inherent sensitivity tradeoff, with the durable fixes for XSS, command injection, and path traversal always structural rather than filtering.

Why AI does not get the structural fix

Parameterized queries work because a database keeps code and data on separate channels, but a language model has one channel where instructions and data are the same natural-language stream, so there is no place to put the boundary. The one defense with a structural fix is exactly the one a language model cannot use, and the jailbreak input grammar (all of human language plus roleplay, translation, and encodings) is effectively infinite.

The tell is in Anthropic's own release notes

Anthropic admits the classifier flags benign coding and debugging more often, the WAF tradeoff stated plainly, and blocked requests fall back to Opus 4.8 which it says is just as capable at the cyber task. So friction lands on everyday coding while the contained capability still sits in the fallback, the export-ban logic reproduced inside one vendor's lineup. A 99% block rate is real and raising attack cost has value, but the shared jailbreak-rating framework and HackerOne program are an admission that new bypasses are expected and will not stop coming.

Fable 5 redeployed globally July 1, 2026; launched June 9 and suspended days later; Mythos 5 restored to Project Glasswing partners.

New safety classifier catches the technique in more than 99% of cases; NIST's Center for AI Standards and Innovation (CAISI) called the safeguards 'extraordinarily strong.'

Blocked requests fall back to Opus 4.8, which Anthropic's own testing says has no unique cyber edge over Fable on the capability in question.

Return terms: Fable counts as half usage through July 7 then moves to pay-per-use credits; enterprise gets nothing included unless credits are on; cloud access on AWS, Google, and Microsoft rolling out 'as quickly as possible.'

References: 'Constitutional Classifiers' (Anthropic Research, Jan 2025) and 'Jailbreaking Large Language Models in Infinitely Many Ways' (arXiv: 2501.10800).

Anthropic, Amazon, Microsoft, Google, and the Glasswing partners are building a shared framework rating each jailbreak by discovery difficulty and extra capability granted; Anthropic stood up a HackerOne program for cyber bug reports.