

REFLEX Shield

User manual: how to operate and stress test it

Built by Dolapo Oyatomi, Reflex Forge | Build 6.2.0 | October 19, 2026

1. What REFLEX Shield is

REFLEX Shield protects payments and AI agent actions from two situations that ordinary fraud checks handle badly. **Hijack** is an attacker using stolen credentials, a compromised device or a manipulated AI agent. **Duress** is a real, authorised person being forced to act. For every request it decides in real time what should happen and shows the reasons, so a security team can see why.

The page has three tabs: **Simulator** (try requests), **Rules** (adjust the demo settings) and **Integrate and test** (self-check, stress test and exports).

2. The four outcomes

Outcome	The person making the request sees	The security team sees
Allowed	The payment goes through normally.	Nothing to do. Logged for audit.
Allowed with alert	It goes through, possibly with an extra check.	An analyst is asked to review.
Decoy activated	A normal success screen. No real funds move.	A silent duress alert. Funds are held while the team responds.
Account frozen	The request is blocked and the session is locked.	Hijack suspected. Credentials frozen and an incident opened.

Every request has an amount and four yes/no warning signs: a new or unlisted destination, a new location or device (or changed instructions), activity outside normal hours, and unusual speed or repetition. The wording adapts to Banking, Crypto, AI agents and Enterprise. The more warning signs and the larger the amount, the stronger the response, and dangerous combinations are always intercepted.

3. How to operate it successfully

1	Choose an industry at the top of the Simulator (Banking, Crypto, AI agents or Enterprise).
2	Start with a preset such as Routine payment, Hijacked agent, Insider under duress or Legitimate but unusual. Or choose an identity, type an amount and tick the warning signs yourself.
3	Choose the ground truth (legitimate or malicious). This is only for the demo, so it can score itself. A real system never knows it, and it never affects a decision.
4	Click Submit request. Read the decision, the score meter, what each side sees, and the list of reasons.
5	Check the Source line. It says whether the live engine or the built-in local engine answered.
6	Use the history table to track results. Export it as CSV or JSON from the Integrate and test tab, and copy the SIEM event line for your monitoring team.

Reading the history column. Correct: a malicious request was intercepted, or a legitimate one was not blocked.

Missed attack: a malicious request got through. **False block:** a legitimate request was blocked.

4. How to stress test it

1	Self-check. Open Integrate and test and click Run self-check. It must report 0 problems.
2	Presets. Run all four presets in all four industries and confirm the results make sense.
3	Amount boundaries. With no warning signs try \$999, \$1,000, \$10,000, \$100,000, \$1,000,000 and \$10,000,000. Risk should rise steadily and never fall.

4	Warning sign combinations. Tick signs one at a time, then in pairs, then all four. The response may only stay the same or get stronger.
5	Bad input. Submit an empty amount and a negative amount. The page must refuse with a message, never approve.
6	Rules tab. Change a setting and repeat a request to see the local result move. Click Reset to defaults when you finish.
7	Stress test button. Run 1,000 cases and note the three percentages.

About the stress test. It runs made-up requests (some malicious, some legitimate) through your current settings. The percentages show how the settings react to that mix. They are **not** real-world detection or false-alarm rates and must not be presented as such. Real rates need a pilot on real, labelled transactions.

5. Using it for a real pilot

1	Start in watch mode. For a few weeks, let it score real transactions without acting on the result, and compare with what your fraud team finds.
2	Fail closed. If the system ever returns anything other than a normal decision, do not approve. Hold the transaction for review.
3	Train staff on the two blocks: freeze means lock and investigate, decoy means keep the person safe and respond quietly.
4	Review weekly. Count missed attacks and false blocks, change one thing at a time, and test again.
5	Do not enter real customer data into the public demo. Pilot setups are arranged separately. Contact Reflex Forge to discuss one.

6. Good to know

REFLEX Shield is a tested, rule-based pilot-ready prototype. It is not a trained model, it has not yet been tested on real transaction data, and it is not a certified production platform. A malicious request that shows no warning signs and a modest amount can pass, and the demo reports that honestly as a Missed attack. Settings and history stay in your own browser.

REFLEX Shield, built by Dolapo Oyatomi, Reflex Forge.