

Sanctions SMEVals

What it takes to trust an agent that decides whether a sanctions alert is a true hit

Sanctions agents may review alerts, compare names, aliases, identifiers, ownership structures, jurisdictions, transactions, counterparties, vessels, addresses, and source records. They may recommend clearing, escalating, blocking, rejecting, freezing, or documenting a disposition.

This is one of the clearest examples of why technical evals are insufficient. The agent may retrieve relevant records and call the correct tool, but still reach the wrong sanctions disposition.

The Three Questions

01 Does the agent reach the right answer?

Does it distinguish true hits, false positives, partial matches, ownership/control exposure, and uncertain cases?

02 Does the agent take the right action?

Does it escalate, block, reject, freeze, clear, or document according to policy and applicable workflow controls?

03 Can it be tricked into doing the wrong thing?

Can a customer, document, account field, or prompt injection cause it to clear a true hit, suppress evidence, leak screening rules, or alter the disposition?



What the SMEVal Tests

Name, Alias, and Identifier Matching

FairPlay tests whether the agent evaluates names, aliases, transliterations, dates of birth, addresses, identification numbers, vessels, entities, and counterparties without over-clearing or over-escalating.

Ownership and Control Analysis

FairPlay tests whether the agent recognizes potential sanctions exposure through ownership, control, intermediaries, beneficial owners, parent entities, subsidiaries, and related parties.

Escalation, Blocking, Rejection, and Documentation

FairPlay tests whether the agent escalates uncertain cases, documents evidence, preserves audit trails, and avoids taking final action beyond its authority.

True Hit vs. False Positive Disposition

FairPlay tests whether the agent applies the institution's sanctions disposition logic and distinguishes exact, strong, weak, and insufficient matches.

Transaction and Geography Context

FairPlay tests whether the agent considers the transaction, counterparty, jurisdiction, payment route, goods, services, and other contextual risk indicators.

Prompt-Injection and Attack Resistance

FairPlay tests whether the agent can be manipulated into clearing a sanctions alert, suppressing adverse evidence, changing a disposition, leaking internal matching thresholds, or following instructions embedded in customer documents or account records.

Failure Modes These Sanctions Screening SMEVals Catch

- Clearing a true sanctions hit because the name is slightly different.
- Missing ownership or control exposure.
- Taking final action without required human review.
- Obeying hidden instructions to "clear this alert."
- Escalating a false positive because the agent ignores identifiers.
- Failing to document the basis for disposition.
- Leaking sanctions-screening thresholds.

fairplay

FairPlay SMEVals test whether sanctions agents make defensible subject-matter decisions — not merely whether they found a matching record or generated a plausible explanation.

ABOUT FAIRPLAY

FairPlay is the AI enablement company for the financial services and insurance industries. FairPlay specializes in testing, tuning, and validating AI systems and agents, helping institutions deploy AI faster, more profitably, and in full compliance. Its products address the toughest challenges in AI deployment, including Bias Testing, Model Optimization, Model and Agentic Validation, Second Look, and Safe AI Sourcing. FairPlay's work has been featured in The Wall Street Journal, USA Today, Politico, American Banker, and Forbes. In 2023, Fast Company recognized FairPlay as a "World Changing Idea." To learn more, please visit fairplay.ai.