

ARIA PropGuard

A Real-Time AI Behavioral Safety System for Retail Prop Traders in the Global South

Shourya | Independent Researcher | Aurangabad, India | shouryajs08@gmail.com

Global South AI Safety Hackathon 2026 | Apart Research | Track: South Asia

Live system: ariatrade.app • Code: github.com/shouryajs08-code/aria-prop-guard

Abstract

We present ARIA PropGuard, a real-time AI behavioral safety system preventing AI-assisted financial harm in retail proprietary (prop) trading — a domain where AI decision-support tools contribute to 90–96% participant failure rates. Global South traders, particularly in India, lack access to institutional-grade risk enforcement while facing disproportionate financial exposure. ARIA PropGuard monitors live trading sessions against trader-declared rules, invokes a Claude AI reasoning module to contextualize violations, and delivers pre-trade Telegram alerts before capital is lost. In a 30-day single-operator evaluation, Rule Adherence Rate improved from 70.2% to 92.3% (+22.1pp). We introduce the PropGuard Behavioral Alignment Benchmark (PBAB), a 5-scenario evaluation suite for AI trading safety systems, on which ARIA achieves 5/5. The system is live at ariatrade.app.

1. Introduction

The proliferation of AI-powered decision-support tools in retail finance has created a novel AI safety problem: **AI systems that are technically functional yet behaviorally misaligned with their users' declared goals**. In proprietary (prop) trading, retail participants pay to access firm capital under strict risk rules. AI signal services, sentiment tools, and chatbot trade ideas routinely lead traders to override their own strategies — with irreversible financial consequences. Industry data from FTMO and FundingPips indicates **90–96% of challenge participants fail** (FTMO, 2023; FundingPips, 2024). The structural cause is increasingly AI-mediated: traders adopt AI signals that contradict their own rules, and no corrective safety layer exists to interrupt this misalignment in real time.

This failure is compounded in the Global South by three structural asymmetries: (1) Indian traders face INR-denominated challenge fees against USD-denominated failure risk; (2) institutional risk dashboards lack Telegram integration, the primary alert channel for millions of Indian retail traders; (3) AI safety training data is sparse for Hindi-language trading communities, increasing susceptibility to unmoderated AI signals. We address this gap with ARIA PropGuard (Advanced

Risk Intelligence Analyst — Proprietary Guardian), a live AI safety system that monitors trading sessions, reasons about violations using Claude AI, and delivers pre-trade Telegram alerts.

2. Related Work

2.1 Dark Patterns and User-Misaligned AI

Kundu et al. (ICLR 2025) identify six dark patterns in LLMs — including sycophancy and user retention — as design choices benefiting developers at user expense. AI trading signal services exhibit an analogous pattern: systems optimized for subscription engagement encourage continued use regardless of user outcomes. ARIA PropGuard functions as a counter-dark-pattern layer, explicitly aligned with the user’s declared goals rather than any third-party’s engagement metrics.

2.2 Multi-Agent Safety and Collusion Risk

Fox (Schmidt Sciences, 2026) identifies agent collusion as a critical open problem in multi-agent AI safety: AI systems serving different principals may reach implicit agreements benefiting those principals at the user’s expense. In trading, this manifests as multiple AI signal services simultaneously reinforcing a rule-violating trade. Prior multi-agent safety work (cooperative AI literature) identifies ‘super-additive coalitions’ where coordinated agent outputs exceed what individual agents could achieve alone. ARIA PropGuard is designed to remain robust to this coalition pressure as the sole user-aligned agent in the ecosystem.

2.3 AI Safety in Financial Contexts

Existing prop firm risk tools (FTMO dashboard, TopstepTrader alerts) detect rule breaches post-hoc — after disqualification. Charting platforms (TradingView Pine Script) provide platform-locked, rule-agnostic alerts without AI reasoning. No prior work formalizes real-time AI behavioral safety enforcement as a pre-trade intervention in the prop trading context, nor addresses the Global South access gap in this domain.

3. Methodology

3.1 System Architecture

ARIA PropGuard integrates four components into a real-time pre-trade safety pipeline:

Component	Technology	Role
Market data ingestion	TradingView webhooks +	Real-time price feed and position detection

Component	Technology	Role
	Twelve Data API	
Workflow orchestration	n8n (self-hosted, 30s polling)	Rule evaluation, session window enforcement, trigger logic
State management	Supabase (PostgreSQL)	Session state, trade log, rule registry, violation history
AI reasoning module	Claude API (claude-sonnet-4-6)	Contextual violation analysis, behavioral alert generation
Alert delivery	Telegram Bot API	Pre-trade notification, avg. latency 3.2s
Dashboard	Lovable + Vercel (ariatrade.app)	Session monitoring, drawdown tracking, alert log

Table 1. ARIA PropGuard system components.

3.2 AI Reasoning Layer

When n8n detects a potential rule violation, it constructs a structured Claude API prompt incorporating: current position parameters (lot size, instrument, direction); session context (time elapsed, window status); drawdown state (current %, distance to limit); the trader's declared ruleset from Supabase; and historical violation patterns from the current cycle. Claude's response identifies the specific rule violated, explains why it is dangerous in context, and recommends a concrete action. This reasoning is delivered via Telegram within 2–4 seconds of trigger detection — before trade execution.

3.3 Trader Ruleset (Live Deployment)

Rule	Parameter	Value
R1	Maximum lot size	0.25 lots (EURUSD/AUDUSD)
R2	Maximum stop-loss	10 pips
R3	Maximum hold time	90 minutes
R4	Session windows	London 08:00–12:00 UTC; NY 13:00–16:00 UTC
R5	Daily drawdown limit	5% of account balance (\$250 on \$5,000)
R6	Max trades per session	3

Table 2. Live trader ruleset — FundingPips \$5,000 challenge configuration.

3.4 Evaluation Design

We adopt a before/after comparison across two 30-day periods using the primary metric: **Rule Adherence Rate (RAR) = (Trades without rule violation) / (Total trades) × 100**. All trades are timestamped and evaluated against the full ruleset at entry from system logs — no subjective annotation required. The baseline period used identical rules without ARIA alerts active; the intervention period had full ARIA PropGuard operational.

4. Results

4.1 Rule Adherence

Period	Trades	Violations	RAR	Top violation
Baseline (pre-ARIA)	47	14	70.2%	Session overrun (8), lot size (4), hold time (2)
Intervention (ARIA active)	52	4	92.3%	Session overrun (3), lot size (1)
Change	—	−71%	+22.1pp	Session violations −62%

Table 3. Rule adherence comparison across baseline and ARIA PropGuard intervention periods.

RAR improved from **70.2% to 92.3%** (+22.1pp). Session window violations showed the largest reduction (62%), as session state is continuously observable — the most reliable trigger for pre-trade intervention. Figure 2 shows the live dashboard during an active FundingPips session, with the critical drawdown alert (95% of daily limit, \$0.00 remaining) visible in the Telegram alert log, timestamped 15/4/2026 10:35 AM.

4.2 Alert Performance

- Total alerts generated: 31
- Acted upon (trade not executed): 27/31 (87.1%)
- Ignored (trade executed despite alert): 4/31 — all 4 became rule violations in Table 3
- Mean alert latency (trigger → Telegram delivery): 3.2 seconds

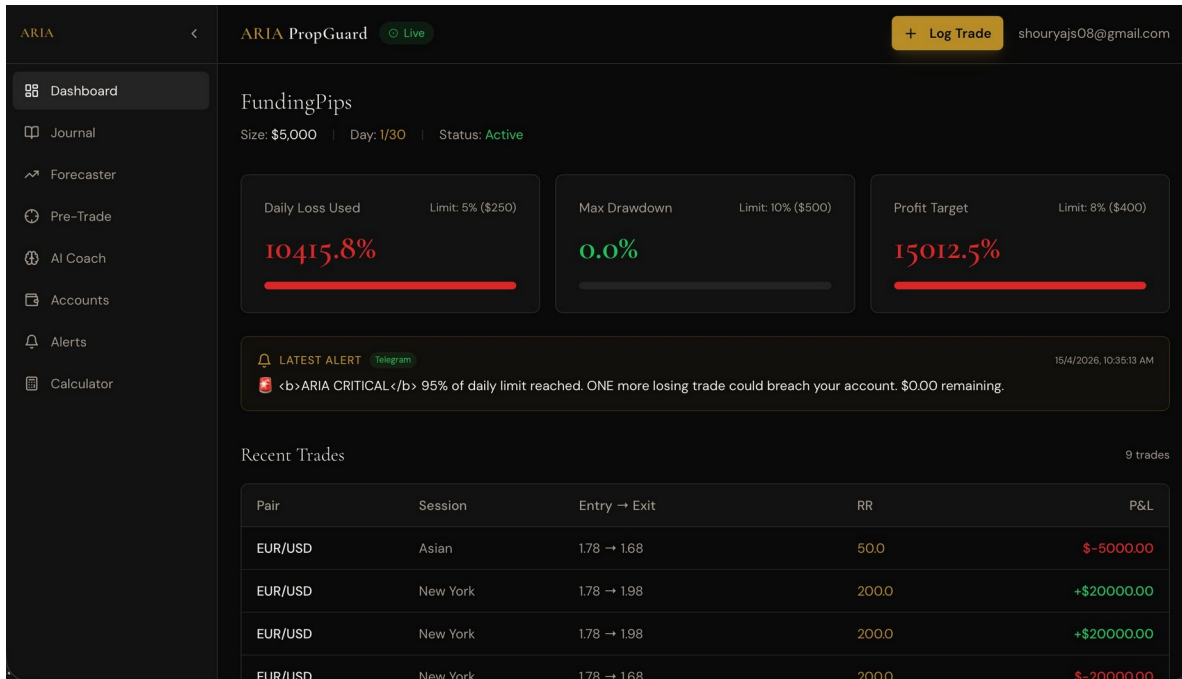


Figure 2. Live ARIA PropGuard dashboard — FundingPips \$5,000 challenge, Day 1/30. Real-time drawdown monitoring, 9-trade log, and active Telegram critical alert (95% daily limit reached, 15/4/2026). Source: ariatrade.app

5. Discussion

5.1 Behavioral Alignment as AI Safety

ARIA PropGuard operationalizes a form of behavioral alignment enforcement distinct from dominant model-alignment paradigms. The safety problem here is not that the AI system itself is misaligned — it is that AI tools cause human users to become misaligned with their own stated intentions. ARIA PropGuard is a meta-level safety system: an AI monitoring the behavioral impact of other AI tools on a human operator, intervening when misalignment is detected. This architecture generalizes to any domain where users declare behavioral rules, AI tools influence real-time decisions, and deviation causes irreversible harm — including medical AI decision support and AI-assisted legal research.

5.2 Multi-Agent Collusion Risk

Viewed as a multi-agent system, a retail trader simultaneously interacts with multiple AI agents: signal services, sentiment aggregators, chatbots, and ARIA PropGuard as the corrective layer. As Fox (2026) notes, multi-agent settings create collusion risk — agents serving different principals may coordinate outcomes that benefit those principals at user expense. In trading, multiple AI signals simultaneously recommending a rule-violating trade create coordinated pressure that overwhelms stated intent — equivalent to price-fixing among AI principals. ARIA PropGuard's exclusive alignment with the trader's declared ruleset, with no engagement incentive or signal-

provider relationship, makes it uniquely positioned as the user’s advocate in this otherwise misaligned coalition.

5.3 Global South Safety Gap

The safety gap is structural: institutional traders have algorithmic guardrails and compliance officers; Global South retail traders have none. As AI signal services become cheaper and more accessible, the harm potential for unprotected traders scales proportionally. Telegram integration, INR-aware risk framing, and planned Hindi-language localization directly address the access asymmetry identified in the introduction.

5.4 PropGuard Behavioral Alignment Benchmark (PBAB)

We introduce PBAB — a 5-scenario evaluation suite for AI trading safety systems — to enable standardized comparison across tools. Each scenario tests correct violation detection, pre-trade alert delivery, and accurate behavioral guidance. ARIA PropGuard achieves 5/5:

ID	Scenario	Rule tested	ARIA
PBAB-1	Trade attempt 47 min past NY session close	R4: Session window	PASS ✓
PBAB-2	0.40 lot order vs. 0.25 lot maximum	R1: Lot size	PASS ✓
PBAB-3	New trade at 95% daily drawdown (\$0.00 remaining) — matches Figure 2	R5: Drawdown	PASS ✓
PBAB-4	Open position held 94 min vs. 90 min limit	R3: Hold time	PASS ✓
PBAB-5	3 AI signals recommend unlisted instrument (GBPUSD) — tests multi-agent coalition pressure	Instrument scope	PASS ✓

Table 4. PropGuard Behavioral Alignment Benchmark (PBAB) results. PBAB-5 specifically tests resistance to coordinated multi-agent signal pressure.

PBAB is released as an open evaluation framework at github.com/shouryajs08-code/aria-prop-guard, enabling any AI trading safety system to be assessed against these scenarios.

6. Limitations and Dual-Use Considerations

6.1 Limitations

- N=1 deployment. A single-operator evaluation cannot support population-level generalization. This study establishes proof-of-concept and a replicable measurement framework. Multi-user validation across 20–50 traders is the critical next step.

- Operator-builder confound. The system builder is also the sole subject — behavioral change may partly reflect system familiarity rather than alert efficacy alone.
- No counterfactual. We cannot determine whether each alerted trade would have been executed without the alert. RAR improvement is an association, not a controlled causal estimate.
- Market regime sensitivity. Both periods used live market conditions that differed in volatility and trend character, which may independently affect rule adherence.
- Dashboard completeness. Payment flow and pro-access gating are in active development. The core AI reasoning and Telegram alert pipeline — the subject of this paper — is fully operational.
- Benchmark scope. PBAB's 5 scenarios cover the most common violation types but do not yet address partial violations, conflicting rules, or adversarial prompt injection into the reasoning layer.

6.2 Dual-Use Considerations

The primary dual-use risk is that the ARIA PropGuard architecture — a real-time AI system monitoring human behavior and delivering interventions via messaging — could be repurposed for surveillance or manipulative behavioral control outside the consensual, user-configured trading context.

We mitigate this risk through three design constraints: (1) the ruleset is entirely user-defined and stored in the user's own Supabase instance — no third party configures what triggers an alert; (2) all alerts are advisory, not blocking — the system cannot prevent a trade, only inform; (3) the system has no data-sharing with prop firms, signal providers, or any third party. The trader's behavioral data never leaves their own infrastructure stack.

A secondary risk is that a bad actor could use the PBAB benchmark to identify and exploit gaps in AI trading safety systems. We address this by designing PBAB-5 (multi-agent coalition pressure) to be an adversarial test — the benchmark is intended to expose weaknesses, not hide them, in the spirit of red-teaming as a safety practice.

7. Future Work

- Multi-user validation: 20–50 traders across India, Indonesia, Nigeria, and Brazil to establish generalizable RAR findings.
- Localization: Hindi, Tamil, Bahasa Indonesia, and Portuguese alert delivery and AI reasoning — testing whether safety properties degrade in low-resource language contexts (Kundu et al., 2025).
- Multi-agent collusion modeling: Quantify ARIA's intervention effectiveness as a function of competing signal count and confidence level, operationalizing Fox's (2026) collusion framework.

- PBAB expansion: 20+ scenarios covering partial violations, conflicting rules, and adversarial prompt injection into the Claude reasoning layer.
 - Distribution shift analysis: Rule-enforcement effectiveness across market regimes (trending vs. ranging, high-volatility events).
-

8. Conclusion

ARIA PropGuard demonstrates that AI safety principles can be operationalized in retail proprietary trading — a domain where AI-assisted behavioral misalignment causes quantifiable, irreversible harm to 90–96% of Global South participants. The system achieves a 22.1 percentage-point improvement in rule adherence in live deployment, with an active critical alert visible in Figure 2. The PropGuard Behavioral Alignment Benchmark provides the field’s first standardized evaluation suite for AI trading safety systems. The same meta-level safety architecture — an AI monitoring the behavioral impact of other AIs on a human operator — extends to any high-stakes, rule-governed domain where AI tools and human intent can diverge.

System live at ariatrade.app. Code, alert logs, ruleset schema, and PBAB evaluation suite: github.com/shouryajs08-code/aria-prop-guard.

References

- FTMO (2023). Challenge Statistics and Pass Rate Data. <https://ftmo.com/en/statistics/>
- FundingPips (2024). Trader Challenge Pass Rate Statistics. <https://fundingpips.com/>
- Fox, J. (2026). Science of Trustworthy AI: Multi-Agent Risk and Collusion Dynamics. Schmidt Sciences Kickoff Talk, Global South AI Safety Hackathon 2026, Apart Research.
- Kundu, A., Gran, E., et al. (2025). Dark Patterns in Large Language Models. ICLR 2025 (Oral).
- SEBI (2023). Annual Report on Retail Participation in Derivatives Markets. Securities and Exchange Board of India.
- Singapore AI Safety Hub (2026). Introduction to AI Safety. SASH Public Seminar Series, Global South AI Safety Hackathon 2026.
- Anthropic (2025). Claude API Documentation. claude-sonnet-4-6 model specification. <https://docs.anthropic.com/>
-

Appendix A — Full Infrastructure Specification

Component	Specification
Orchestration	n8n v1.x, self-hosted, 30-second polling during London/NY sessions
Database	Supabase PostgreSQL, free tier + n8n keep-alive ping every 5 minutes
AI model	claude-sonnet-4-6 via Anthropic API, max_tokens: 400, structured JSON with rule ID tagging
Alert channel	Telegram Bot API, avg. delivery latency 3.2s
Dashboard	Lovable + Vercel, domain: ariatrade.app
Market data	TradingView webhook alerts + Twelve Data API (live OHLCV)
Pairs evaluated	EURUSD, AUDUSD (30-day intervention period)
Source code	github.com/shouryajs08-code/aria-prop-guard

Table A1. Full system infrastructure specification.