AceGuard (SN87) Whitepaper

Behavioral bot detection for competitive systems, starting with online poker

Abstract

AceGuard is a Bittensor subnet designed to produce reliable, probabilistic signals for detecting automated agents that impersonate humans in competitive environments. The initial domain is online poker because it concentrates incentives, adversarial behavior, and rich sequential decision-making into a measurable setting. AceGuard's miners produce probabilities. Validators score miners using controlled, versioned datasets with ground truth and apply asymmetric penalties that strongly discourage false positives. The long-term goal is to create a generalizable behavioral security layer for any competitive platform where trust is fragile and incentives attract automation.

1. Problem

1.1 The bot problem is structural, not temporary

Competitive platforms that offer financial rewards, status, or rankings inevitably attract bot automation and cheaters. As models and tooling advance, bots are increasingly capable of imitating human behavior, rendering naive detection methods fragile and ineffective. Bot detection in this context is not a problem that can be solved once, but an ongoing arms race requiring continuous adaptation and iteration.

Online poker amplifies this problem. The game is inherently adversarial, the incentives are immediate and monetary, and the ecosystem generates vast amounts of interaction data.

Despite this, detection efforts remain fragmented and proprietary, siloed within individual poker operators. This fragmentation creates a clear opportunity for a specialized business focused on the cybersecurity and integrity of competitive online poker.

1.2 The key risk: false positives

Mistaking a human for a bot is reputationally and ethically costly. In competitive communities, false accusations are often worse than missed detections. Any scalable solution must be conservative by design, express uncertainty, and avoid acting as judge and jury.

2. Goals and Non-goals

2.1 Goals

AceGuard aims to:

- Detect automation with high confidence by generating probabilistic bot-likelihood signals from behavioral data and metadata, backed by rigorous evidence.
- Maintain strict evaluation discipline through ground truth datasets, versioning, reproducible benchmarks, and a conservative bias that penalizes false positives.
- Build for extensibility by developing methods that generalize beyond poker to other competitive, adversarial domains over time.

2.2 Non-goals

AceGuard does not:

- Ban accounts, remove players, or take enforcement actions.
- Output definitive labels as a final judgment.
- Depend on privileged platform telemetry as a hard requirement.

AceGuard is an intelligence layer: it generates risk signals that downstream systems may use alongside other evidence.

3. Why Poker First

Poker is an unusually strong testbed for behavioral detection:

- High incentive pressure: direct monetary outcomes.
- Sequential structure: long action histories create rich behavioral traces.
- Adversarial intent: bots actively try to appear human.
- Natural evaluation: outcomes can be measured, and simulation is feasible.

Poker is not the end goal. It is a stress test that forces the system to learn meaningful behavioral signals rather than superficial cues.

4. Subnet Overview (Bittensor)

4.1 Roles

- Miners: submit models that map a sample (e.g., set of hands or sessions) to a probabilistic bot-likelihood score.
- Validators: run evaluations on versioned datasets with ground truth and compute scores.
- Subnet: coordinates incentives so the best-performing miners earn rewards.

5. Data and Ground Truth Strategy

5.1 Two primary data sources

AceGuard uses:

1. Bot-generated data from controlled sandboxes where agent identities are known.
2. Human data from curated hand histories and real play logs where provenance is known and labeling policies are explicit.

The initial benchmark prioritizes tightly controlled ground truth. Over time, AceGuard will operate its own no-real-money platform to collect high-fidelity, real-time behavioral data directly from players.

6. Evaluation Framework

6.1 Scoring objectives

Validators compute miner scores primarily based on:

- **F1 score**: overall detection effectiveness, balancing precision and recall.
- **False positive rate**: heavily weighted to penalize incorrect human classifications.
- **Average precision**: quality of ranking bot samples above human samples across thresholds.

7. Miner Interface

7.1 Input schema

Miners receive a standardized JSON representation of hands and metadata. The schema is stable within a dataset version. Fields may include:

- Hand actions with positions and bet sizes
- Game format parameters (blinds, stacks, table size)
- Optional timing features (when available and permitted)
- Player identifiers anonymized or mapped to stable IDs within a sample

Miners submit a bot-probability score which will get evaluated.

8. Roadmap: The Path to Decentralized Digital Trust

V0 — Foundation: Controlled Detection

Establish a rigorous baseline for reliable automation detection in poker.

- Poker bot detection in a fully controlled environment
- Basic bot pattern and behavioral signal identification
- Ground-truth dataset generation with clear provenance
- Risk scoring system with evidence-backed outputs
- Validator metrics framework with conservative false-positive penalties

V1 — Advanced Detection & API

Expand behavioral realism and make detection consumable by platforms.

- Sophisticated bot detection (timing, adaptation, multi-factor behavior)
- Diverse human behavioral modeling to reduce false positives
- Open API for external platform integration
- Detection dashboard with player-level insights and explanations
- Early-warning and preemptive detection capabilities

V2 — Commercial Scale

Operationalize AceGuard for production and operator workflows.

- Pilot integrations with live platforms
- Comprehensive API and SDK suite
- Risk and evidence dashboards for operators and auditors
- Automated action and response workflows (non-enforcement)
- Expansion across multiple games and competitive environments

V3 — Multi-Platform Expansion *(Major Milestone)*

Move from offline analysis to real-time, platform-agnostic detection.

- Real-time, in-game detection pipelines
- Platform-agnostic behavioral detection framework
- Anti-overfitting evaluation using unseen bot families
- Score calibration and reliability guarantees
- Extended support for multiple game formats and structures

V4 — Global Trust Infrastructure

Evolve into a universal layer for behavioral verification.

- Cross-platform behavioral identity and analysis
- Integration with social trading, marketplaces, and incentive systems
- Automated dataset evolution and continuous benchmarking
- Universal behavior validation under economic incentives

AceGuard remains anchored to measurable behavior, versioned ground truth, and conservative decisioning.

9. What Success Looks Like

AceGuard succeeds if it becomes:

- **A trusted, independent verification layer** that provides accurate, high-quality signals platforms can integrate into their broader security and integrity stack.
- **An adaptive system that continuously improves** as adversarial automation evolves over time.

Success is measurable improvement over SOTA techniques with minimized false positives.

10. Conclusion

AceGuard is building a behavioral security layer for competitive systems, starting with online poker as the most demanding environment. The subnet is designed around realism, measurement, reproducibility, and conservative outputs. By combining controlled ground truth, versioned datasets, and incentive-aligned scoring that punishes false positives, AceGuard aims to push bot detection toward rigorous, trustworthy signals rather than fragile rules or overconfident judgments.