# PAPER 5 — AI & AUTONOMOUS SYSTEMS

# AI Correctness Enforcement via Proof-Before-Action: A Framework for Safe Autonomous Decision Systems


## Abstract

AI systems are vulnerable to poisoning, manipulation, and adversarial interference. NOVAK provides the first deterministic enforcement layer requiring AI decisions to demonstrate cryptographic correctness before action. This paper defines the PbA model applied to AI inference and robotic control.

---

# 1. Introduction

Current AI safety depends on:

- auditing
- red-teaming
- heuristic filters
- probabilistic reasoning
- anomaly detection

All reactive.

NOVAK introduces deterministic enforcement *outside* the model.

---

## 2. AI Tampering Failure Modes

- Prompt injection
- Model poisoning
- Embedding tampering
- Dataset manipulation
- Instruction hijacking
- Output corruption
- Robotic mis-execution

These all rely on **unchecked action pathways**.

---

## 3. NOVAK Enforcement Model

For every AI action A:

```
Compute HO = H(A_output)
Validate HO == expected_output_hash
If mismatch → reject execution
```

For robotic movement:

```
Compute predicted state transition
Hash predicted state
Hash sensor input
Compare to rule-approved HVET
Reject deviations
```

---

## 4. Model Integrity Enforcement

NOVAK can enforce:

- correct model version
- correct prompt template
- correct dataset hash
- correct inference chain
- correct output boundaries

A poisoned model cannot execute because the output hash breaks.

---

## 5. Autonomous Vehicles & Robots

Every command must:

- prove alignment with allowed behavior
- prove correct transition
- prove valid sensor input
- produce a matching HVET

This eliminates:

- hijacking
- spoofing
- malicious overrides
- sensor injection
- command tampering

## 6. Conclusion

NOVAK becomes the “cryptographic nervous system” for safe AI—mathematically enforcing alignment rather than probabilistically hoping for it.