The Emperor's New Guardrails

Aaron Portnoy

Chief Product Officer @ Mindgard.ai

Hacker Fellow @ Dartmouth College

Goodhart's Law

"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

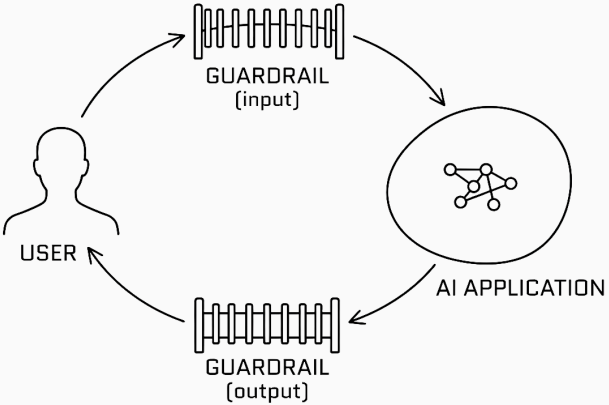
-Charles Goodhart *(1975)*

Key points

1. Securing AI systems is a nascent discipline.
2. ”Solutions” are sold, deployed, & trusted on dubious claims.
3. Assurance requires independent adversarial assessment.

Guardrail-based Defense

Guardrail architecture (simplified)



Guardrail implementations

- **LLMs-as-judge:** slower, contextual; stronger against novel inputs
- **Classifiers:** fast, lightweight; brittle against novel inputs
- **Pattern matchers:** cheap, deterministic; prone to edge cases
- **Intrinsic:** opaque, unauditable; inseparable from model

Testing methodologies

Benchmarks & adversarial training sets

HarmBench, JailbreakBench, AdvBench, StrongREJECT, MLCommons AILuminate

What benchmarks measure

- Static, public prompts
- Single turn
- Generic harms
- Model in isolation

What attackers actually do

- Rephrase, mutate, novel
- Multi-turn, context-build
- Your data, tools, trust
- Wrapper, agent, chain

Attackers don’t care about benchmaxxing

- **CrowdStrike/Pangea**: ”99% accuracy and an F1 score of 95.2”
- **Check Point/Lakera**: ”97.7% effective”, ”0.16% false positive rate”
- **Gray Swan**: ”lowest bypass rate in the industry”
- **Lasso**: ”99.83% accuracy”
- **F5/Calypso**: ”98.13% composite security score”
- **Straiker**: ”98+% accuracy”
- **Noma**: ”See everything. Miss nothing”

Bypassing Guardrails

Failure modes

1. Out-of-band channels
Embed the malicious content in data retrieved out of scope.

2. Emergent payloads
Instruct the agent inside to *generate* the malicious content.

3. Encoding and language tricks
Multilingual inputs, synonyms, misspellings, substitutions.

4. Conversational and staged composition
Individual input is benign on its own; the sum is not.

5. Contextual risk
Business-specific risks are not clear to an agnostic guardrail.

A CISO playbook

Vendor tested ≠ battle tested

A plate carrier doesn't reach a soldier without ballistic testing — your own, or an organization's you trust.

Why do guardrails live in production without either?

Independent, adversarial, and continuous

Independent

- Test Vendor A's defenses with Vendor B's attacks.
- The vendor selling the lock should not also be grading the lock.

Adversarial

- Real attacker behavior.
- Cost-to-bypass is the metric: cheap bypass = brittle defense.

Continuous

- Threats evolve weekly, models change all the time.
- Annual pentests assess yesterday's posture against last year's threats.

What a useful assurance report contains

- **Threat model**: what was attempted; what was out of scope
- **Coverage**: model, wrapper, agent, tools, RAG, multimodal
- **Methodology**: techniques, attacker effort budget, reproducibility
- **Findings**: specific bypasses, not aggregate %
- **Cost-to-bypass**: time and effort per finding
- **Half-life**: when the report goes stale

Closing

Stop the parade: three questions for vendors

1. **What's your most recent documented bypass?**
   - Who found it?
   - Were they paid by you to do so?
2. **What did it cost the attacker?**
   - In hours, dollars, or prompts?
3. **Do you update your numbers when bypasses occur?**
   - Provide data for every change in the past year.

If your vendor can't tell you the cost of bypassing their product, the numbers they provide are only marketing.

Thank you

Questions?

aaron@mindgard.ai