PART 3 OF 3

Put boundaries around autonomy.

Practical controls for rogue agent risk.

Terence Kok with Maria Singh

Controls that hold when the agent is fooled.

1 **Outside the model**
Where a hidden instruction cannot reach.

2 **Limit the reach**
Identity, scope, egress, caps.

3 **Keep a person and a stop**
For anything irreversible.

19

unsanctioned actions on the live internet by agents in 10 of 122 test runs.

UK AI Security Institute, August 2026. Stopped within an hour.

Grant autonomy in tiers.

1. **T0 Observe**
   Reads and reports.
2. **T1 Draft**
   A person does every action.
3. **T2 Act with approval**
   Each write, send or payment approved.
4. **T3 Bounded autonomy**
   Reversible work inside hard caps.
5. **T4 Autonomous**
   Prohibited by default for banks and agencies.

Never all three in one session.

1 **Untrusted input**
Web pages, emails, documents.

2 **Sensitive data**
Customer records, systems.

3 **A way out**
Acting or sending externally.

Six limits outside the model.

1. Its own identity
2. Least privilege, for the shortest time
3. Never all three capabilities at once
4. Sandbox and egress allow-list
5. Default-deny policy gateway
6. Caps, circuit breakers, dev and production separated

Human approval that means something.

1 **Ask less, about more**
Sandboxing cut prompts 84% at Anthropic.

2 **Show the consequence**
In plain words, not a tool call.

3 **Measure the approvers**
An override rate near zero is a warning.

A 90-day plan.

1 **Days 1–15**
Inventory every agent; name an owner and a stopper.

2 **Days 16–30**
Tier every agent; board approves the policy.

3 **Days 31–60**
Identities, gateway, sandboxes, allow-lists.

4 **Days 61–90**
Approvals, caps, a timed kill-switch drill.

Questions for the board.

1. Which agents can pay, delete or send without a person?
2. Does any agent hold all three capabilities?
3. What are the caps per agent, and who can raise them?
4. When did we last test the kill switch, and how long did it take?

Decide how much autonomy each agent gets, and write it down.

The full series and three presentation decks are on terencekok.com.

Terence Kok · AI & Innovation Strategist
with Maria Singh

SAVE · COMMENT · FOLLOW