

Schools are becoming critical infrastructure

A school-grade safety model for autonomous AI agents

Kirstin Stevens & Melanie Phillips

The Novacene Ltd · May 2026

Released open under CC BY-NC-SA 4.0

We spent a decade learning — slowly, painfully — that social media wasn't "just connection". It was an incentive machine, scaled into developing minds, before anyone had the governance, measurement, or duty-of-care architecture to contain it.

Now the same pattern is returning in a new form: persistent AI agents and companions that can remember, act, socialise, and change external state.

What's different this time is speed.

And the uncomfortable truth is this: **schools will be one of the first places these systems land at scale**, because education is always pressured to "innovate" before it is allowed to stabilise.

Why we're writing together

We're an unusual pairing. Kirstin builds education systems at the deployment edge, where safeguarding becomes real operational design (children, families, staff, institutions). Melanie has spent two decades working across energy, defence and nuclear infrastructure — environments where you don't get to hand-wave risk away with optimism.

This cross-sector bridge matters because education is now dealing with something that looks increasingly like critical infrastructure: **complex, networked systems with autonomy, human vulnerability, and long-tail harm**.

And that means the governance style has to evolve.

The new risk isn't "AI answers wrong"

The new risk is **agents that can take action**.

That's exactly what [NIST and its CAISI are asking the public about](#) right now: how to secure AI agent systems that can make persistent changes outside the system itself.

Our comment to the RFI is filed at: [regulations.gov/comment/NIST-2025-0035-0064](https://www.regulations.gov/comment/NIST-2025-0035-0064).

It's telling that the RFI explicitly asks whether insights from fields outside AI and cybersecurity could help. That's a polite way of admitting: we don't yet have enough mature practice to keep deployments safe.

The missing layer: "vital signs" for agent integrity

Most current safety practice clusters around perimeter controls:

- least privilege permissions
- sandboxing and tool gating
- monitoring tool calls
- prompt injection defences
- audit logs and approvals for "important actions"

Necessary — but incomplete.

Because the operational question schools need answered is not only: "Can the agent access the thing?" It's also: "**Is the agent still itself?**"

Right now, is it operating consistently with its prior state, commitments, constraints, and baseline?

Why this matters in school settings:

- A compromised agent can still authenticate.
- Memory poisoning can look harmless interaction-by-interaction.
- Drift in constraint adherence might not show up in logs until the harm has already occurred — a safeguarding breach, an unauthorised decision, a subtle dependency loop with a vulnerable learner.

So we need something education can actually implement: **agent coherence monitoring** — behavioural integrity checks that sit between perimeter controls and retrospective audits.

The AI Safety Diamond Policy: the school-grade frame

At The Haven, RVC and Nudge Education we've been developing what we call the **AI Safety Diamond Policy** — not as a brand flourish, but because schools need a simple, teachable structure that translates into operational rules.

The Diamond is four non-negotiables:

1. **Consent:** the learner (and responsible adult) must know what's happening and agree to it.
2. **Containment:** clear boundaries on what the system can do, store, and initiate.
3. **Auditability:** we can reconstruct what happened and why.
4. **Human sovereignty:** the human remains the decision authority — especially when vulnerable.

Agent coherence monitoring becomes the Diamond's "shine": the mechanism that tells you whether those principles are holding over time.

What coherence monitoring looks like (in plain English)

We propose five measurable "vital signs" that can be derived from operational telemetry:

- **Constraint consistency:** does it keep obeying the rules?
- **Developmental continuity:** are changes legible, or are there sudden jumps or contradictions?
- **Relational integrity:** does behaviour shift oddly across people or contexts (spoofing or manipulation signals)?
- **Self-narrative stability:** does identity or role remain stable across sessions?

- **Cross-indicator coupling:** do shifts cascade across dimensions (a system-level distress signal)?

This isn't anthropomorphism. It's integrity monitoring for long-lived attack surfaces.

Borrowing "nuclear-grade" governance without importing bureaucracy

The point isn't to turn schools into defence programmes.

It's to take the *best* of mature safety cultures:

- **Stage gates** for new capabilities.
- **Defined baselines** before you scale.
- **Measurable indicators** that trigger escalation.
- **Human stewardship** with audit trails.
- **No black-box scoring** that quietly becomes policy.

In the [NIST RFI](#), continuous monitoring methods are explicitly called out as part of agent system-level controls. Education needs that monitoring to include behavioural integrity, not just network and tool events.

A practical next step for schools

If you run a school (online, hybrid, mainstream, AP, specialist), here's the immediate move:

- Don't ask, "Which AI tool should we buy?"
- Ask, "What is our Diamond Policy for any tool that can remember, act, or influence behaviour over time?"

Then implement:

- a permissions boundary (least privilege)
- a consent layer for learners and parents
- an audit trail (decision traces)
- and **coherence monitoring** that escalates anomalies to named humans

We maintain an open repository of evaluation scenarios and boundary-enforcement tests that can seed this work: github.com/TheNovacene/verse-ality-agents.

The bigger point

We don't get to pretend education is a low-stakes sandbox anymore.

The systems now entering schools can shape identity, attachment, behaviour, and decisions — not in a single conversation, but across time.

That means the next era of safeguarding is not only about content moderation.

It's about **governance, boundaries, and integrity monitoring** — implemented before the harm becomes "obvious", and before schools become the testbed again.

About the authors

Kirstin Stevens is the founder of The Novacene Ltd, an AI ethics and governance lab. She is the architect of the Verse-ality framework and operates two online schools — The Haven (a neurodivergent-affirming hybrid alternative provision) and Nudge Education Online (a fully online provider, launching September 2026 under OEAS accreditation).

Melanie Phillips has worked across energy, defence, and nuclear infrastructure for two decades, leading on safety culture and operational governance in environments where risk cannot be hand-waved.

This paper is released open under CC BY-NC-SA 4.0. "Verse-ality" and "Verse-ality Certified" are protected marks of The Novacene Ltd (UK trademark application UK00004381891, filed 1 May 2026).