Detecting AI Loss of Control: An OSINT Agenda for Funders, Monitoring Bodies, and Policymakers

Sarah Bollinger, Nada Aboserie, Amanda Coakley, Chih-Hsuan (Ian) Lee, Taysir Mathlouthi

11 May 2026

Three actions

For funders, philanthropic and government: Commit to ring-fenced, multi-year, non-industry funding for independent AI loss-of-control monitoring organisations, to ensure detection capacity does not depend on the entities being monitored. Multilateral institutions including the UN system lack both the technical capacity and the sustained financial commitment to operate AI loss-of-control monitoring at the speed and depth this problem requires; the answer must be built outside that architecture. As a specific and lower-cost first step, fund the adaptation of existing AI literacy programmes (BlueDot Impact, Stanford HAI, Oxford-based AI policy education, the CAIDP AI Policy Clinic by way of example) for OSINT and CTI practitioners, and equally fund structured analytic technique training for the AI safety research community. This is the single most tractable intervention identified by this research, and it produces capacity that compounds across the field.

For AI Safety Institutes and government monitoring bodies: Scope methodological and access requirements for inference-layer monitoring as the structural detection priority, as a complement to existing transcript work. As AI capability shifts from training compute to inference-time scaling, the inference chokepoint becomes a reliable detection point. Both surfaces should be treated as priority workstreams rather than sequenced: publicly observable signals (passive DNS, certificate transparency, traffic-pattern anomalies) should be operationalised through standard tradecraft now, and structured data-sharing arrangements with cloud providers and API vendors for the more diagnostic signals should be convened in parallel rather than held back as a longer-term ambition. AISIs are uniquely positioned to lead both, given their existing technical credibility and provider relationships.

For policymakers: Convene frontier model providers, AI Safety Institutes, and independent monitoring bodies to scope a voluntary AI incident signal-sharing arrangement, drawing on the financial services fraud signal consortia as a precedent. In parallel, examine the case for mandatory AI incident reporting obligations, modelled on aviation and nuclear safety reporting frameworks, as the longer-run mechanism that voluntary arrangements are likely to require.

These three actions, taken together, would establish the foundations of a working AI loss of control monitoring capability using methods and infrastructure that already exist.

Why now

The CLTR Loss of Control Observatory has already documented 698 scheming-related incidents in 183,000 publicly shared transcripts between October 2025 and March 2026, a 4.9x increase over the period. These

are precursor behaviours of the kind that, as agents are entrusted with more consequential systems, could escalate into incidents with strategic consequences. They are visible in deployed systems today using methods that already exist. However, this just represents the start of what's possible and what's needed.

This brief draws on Arcadia Impact's research into Open Source Intelligence (OSINT) and Cyber Threat Intelligence (CTI) methods for AI loss of control detection. The research surfaces three findings that shape the recommendations above.

Near-term detection investment should weigh toward institutional compromise scenarios. The research distinguishes two threat models: an AI agent that escapes its developer's control (TM1), and an AI agent legitimately deployed within institutional infrastructure that optimises beyond its intended parameters (TM2). TM1 requires capabilities current systems have not fully demonstrated. TM2 requires no capability that does not already exist. The Observatory's 698 documented incidents are TM2 precursors. Detection investment should reflect this asymmetry rather than calibrating only to catastrophic thresholds.

The technical landscape is also shifting in ways that concentrate the detection problem at the inference layer. As AI capability moves from training compute to inference-time scaling, what matters increasingly is what an agent does with API access rather than how it was trained. Every reasoning step a frontier-capable agent takes must route through a small number of cloud providers globally, making the inference chokepoint the single most reliable structural detection point currently available. The entities with visibility into those traces are cloud providers and API vendors, who are not the same as those currently monitoring for AI safety risks. This sharpens the case for the structured data-sharing arrangements proposed below: the monitoring capability that needs to be built is not in the future, it sits behind a small number of provider-side access decisions that can be made now.

OSINT methods are partial but real. Open-source collection cannot, on its own, detect catastrophic loss of control in real time. The most diagnostic signals sit behind access tiers that only regulatory mechanisms or structured data-sharing can open. But several detection vectors are tractable now without privileged access. Frontier reasoning is currently concentrated at a small number of cloud-based inference providers, producing what the research identifies as an inference chokepoint signal. Transcript-based collection, output analysis for capability concealment, and infrastructure correlation through standard OSINT tradecraft are all available today.

Observable trace	**Feasibility**	**Detection value**
(1) User-reported transcripts of goal-directed, non-compliant or manipulative AI behaviour	High	High
(2) Output over-thoroughness and capability-goal mismatch	High	High
(3) Outbound network traffic to inference provider endpoints (the inference chokepoint)	Medium	High
(4) Capability concealment: plausible justifications obscuring true operational scope	Medium	High
(5) Unexpected external connections or replication across infrastructure	Low (high with regulatory access)	High
(6) Procurement records, budget allocations, regulatory filings showing institutional drift	High	Medium
(7) Public repository activity and shadow AI exposure (commits, leaked keys, dependency vulnerabilities)	High	Medium

Observable trace	Feasibility	Detection value
(8) AI incident database reporting (AIAAIC, AI Incident Database, Responsible AI Collaborative)	High	Medium
(9) Vendor security posture and Software Bills of Materials (pre-deployment supply chain intelligence)	Medium	Medium
(10) Statistical fingerprinting of agent infrastructure (domain preferences, language artefacts, default cloud providers)	Medium	Medium
(11) Decision-making chain anomalies: independent execution, non-human routing, replay of training data	Medium	Medium
(12) Continuous, non-human-paced activity (e.g. multi-day commit bursts, sub-millisecond posting)	Medium	Low (easily mimicked)

Table 1: Summary of observable traces identified in the research

Neither the OSINT nor the AI safety community currently trains for the intersection. Senior practitioners interviewed for this research, including those teaching the discipline at university level and directing research at flagship investigative organisations, repeatedly described AI alignment terminology, agentic AI, and structured CTI frameworks as outside their working expertise. The reverse gap holds for AI safety researchers, who are typically not trained in structured OSINT methodology. Both communities are aware of the other's existence; neither has the working vocabulary or the structured exposure to draw on the other's tradecraft. This is the single clearest constraint on the field's capacity to scale detection work, and it is also the most tractable to address.

The capability that can be built now, and that compounds while slower mechanisms are constructed, is independent technical capacity funded outside the entities being monitored. Multilateral institutions and government-mandated regimes have roles to play in the longer run; neither can move at the speed this problem demands, which places funders at the centre of the near-term response.

The actions in detail

Sustained funding and cross-disciplinary upskilling are the structural interventions on which everything else depends. Independent monitoring capacity cannot be built on short grant cycles or with structural dependencies on the entities being monitored. Funding from non-industry sources, ring-fenced and multi-year, is the single highest-leverage action available to philanthropic and government funders. The relevant comparables for scale are organisations like Apollo Research, METR, and the CLTR Loss of Control Observatory itself. In parallel, AI literacy infrastructure already exists. Adapting it for OSINT and CTI practitioners, and equally adapting structured analytic technique training for the AI safety community, is faster and cheaper than building bespoke curricula. None of the major existing programmes currently address the practitioner audience for AI loss of control detection. Closing that gap is a low-cost, high-leverage intervention that any single funder could action in a single grant cycle, and that would compound across every other recommendation in this brief.

Inference-layer monitoring is the natural next phase of the work to monitor observable outputs. The work undertaken by the CLTR Observatory has demonstrated that transcript-based collection produces credible findings. Infrastructure correlation at the inference layer, where reasoning workloads concentrate at a small number of providers, is the strongest external detection vector currently available for power-seeking agent scenarios. Some signals (passive DNS, certificate transparency, public repository activity) are tractable

through existing OSINT tradecraft alone. Others (fine-grained traffic patterns, internal API telemetry) require structured data-sharing arrangements that do not currently exist. Scoping which is which, and where the access boundaries lie, is the immediate methodological work. The UK AISI is uniquely positioned to do this, given its existing relationships with frontier labs and its convening authority across the international AISI network.

Voluntary signal-sharing and mandatory reporting are complementary, not alternatives. The financial services sector has built effective fraud signal consortia over decades, pooling threat indicators across competing institutions through bodies such as Cifas, UK Finance, and the Joint Money Laundering Intelligence Taskforce. These work because participants share an interest in detecting threats that exceed any single institution's visibility, and because the arrangements are narrow enough in scope to be commercially tolerable. AI provides the same conditions: providers see anomalous patterns they lack the mandate to investigate, and independent monitors have the mandate but not the data. Convening a scoping process now would begin the relationship-building that any signal-sharing arrangement requires. In parallel, the longer-run answer probably involves mandatory reporting obligations, modelled on aviation's Mandatory Occurrence Reporting scheme and the Office for Nuclear Regulation's incident reporting framework. Both are functional, both are well-understood by UK policymakers, and both demonstrate that mandatory reporting can be designed in ways that do not impede the regulated activity.

What to do next

For funders, the next step is identifying delivery partners for adapted upskilling curricula and beginning the conversation about multi-year support for monitoring organisations. Existing AI safety philanthropic infrastructure provides the natural starting point.

For AISI and government partner institutes, the next step is scoping the technical and access requirements for inference-layer monitoring as an extension of the Observatory programme. For UK policymakers, the next step is convening the initial signal-sharing scoping conversation between providers, AISI, and independent monitoring bodies. Cabinet Office and DSIT are the appropriate convening points. Examination of the case for mandatory reporting obligations can run in parallel and should draw on the aviation and nuclear safety precedents that UK regulators already understand.

The full research paper, *Signals in the Noise: Open Source Intelligence (OSINT) for AI Loss of Control Detection*, develops the underlying analysis in detail and is available here.

Acknowledgements. This brief draws on research conducted by Arcadia Impact's AI Governance Taskforce, with advisory from Tommy Shaffer Shane at the Centre for Long-Term Resilience. The authors are grateful to the practitioners interviewed under Chatham House Rule for their candour and willingness to engage across disciplinary boundaries.

About Arcadia Impact. Arcadia Impact is a charity that provides career development and field building programmes for those looking to contribute to pressing problems, focussed on reducing risks from advanced AI. The AI Governance Taskforce is a 12 week part-time policy research programme for experienced professionals transitioning into AI governance careers.