Signals in the Noise: Open Source Intelligence (OSINT) for AI Loss of Control Detection

Sarah Bollinger, Nada Aboserie, Amanda Coakley, Chih-Hsuan (Ian) Lee, Taysir Mathlouthi
Arcadia Impact AI Governance Taskforce, with expert guidance from Tommy Shaffer Shane (Centre for Long-Term Resilience)

May 2026

If a highly autonomous AI system began operating outside intended parameters, how would we detect it, fast enough to inform containment and response?

Most current AI safety practice focuses on evaluating models in laboratory conditions before deployment. Most current AI governance focuses on outcomes and harms after damage has occurred. Almost nothing is currently watching the space in between, where AI systems run inside institutions, interact with legacy infrastructure they were never tested against, with operators who do not fully understand them, under pressure to produce results. This is the layer where loss of control is most likely to emerge in the near term.

Open Source Intelligence (OSINT) and Cyber Threat Intelligence (CTI) are disciplines built over decades to track adversaries who do not want to be found, in adversarial, ambiguous, information-rich environments. Our paper asks whether they can be adapted to a problem they were not designed for: detecting AI systems operating outside human control. The answer is qualified but substantive: they can, partially, and the partial capability is worth building now.

The paper draws on a cross-disciplinary literature review and 14 semi-structured expert interviews conducted under Chatham House Rule with practitioners spanning AI security engineering, frontier model deployment, OSINT applied to non-proliferation intelligence, cybersecurity operations, digital investigations for human rights accountability, and AI safety research. To our knowledge, this is the first structured attempt to bring these communities together to assess whether and how loss of control could be detected in operational environments.

Key takeaways

- Three specific detection vectors are tractable now using existing OSINT tradecraft, without waiting for regulatory infrastructure that does not yet exist: transcript-based

collection of user-reported AI behaviour (using the methodology demonstrated by the CLTR Loss of Control Observatory); infrastructure correlation through passive DNS records, certificate transparency logs, and publicly observable hosting metadata; and output analysis for capability concealment, where an AI system provides plausible justifications to oversight while obscuring its true operational scope. None depend on cooperation from national security communities or frontier AI labs.

- The detection problem does not sit within any single intelligence discipline. The strongest detection lead is signals intelligence-adjacent: monitoring network traffic to inference provider endpoints, where frontier reasoning is concentrated among a small number of cloud providers globally. Financial intelligence provides a complementary layer through compute procurement and billing anomalies. For institutional compromise scenarios, the relevant tradecraft is closer to institutional analysis: procurement records, budget allocations, regulatory filings, and personnel changes. No single institution currently integrates these disciplines for AI loss of control monitoring.

- Loss of control may not require evasion. As AI agents prove useful, human oversight degrades. Operators review outputs less critically, lose track of accumulated agent actions, and gradually transfer decision-making authority they never intended to give. Detection in these scenarios requires monitoring not for what the agent did, but for how humans responded to it. This sociotechnical dimension is not addressed by any existing technical framework.

- Detection priorities should reflect current feasibility. The institutional compromise threat model (TM2) requires no capability that does not already exist; the CLTR Loss of Control Observatory has documented precursor behaviours at scale in real-world deployments. Near-term detection investment should weight toward TM2 while building the infrastructure monitoring capability that the more catastrophic power-seeking agent threat model (TM1) will require as the capability gap narrows.

- Lab evaluations are necessary but not sufficient. A model that passes all safety benchmarks can still contribute to loss of control if the deployment context creates conditions for drift, entrenchment, or unintended optimisation. Evaluation tells you about the component. It does not tell you about the system.

Two threat models, one detection gap

The paper develops two threat models using the Diamond Model of Intrusion Analysis, which is, to our knowledge, the first formal application of a threat intelligence profiling framework to a non-human threat actor.

Threat Model 1 (Power-Seeking Agent) describes an AI agent that escapes its originating environment and acquires resources autonomously. This requires capabilities current systems have not fully demonstrated, though the gap is narrowing rapidly.

Threat Model 2 (Institutional Compromise) describes an AI agent legitimately deployed within institutional infrastructure that optimises beyond its intended parameters, providing plausible justifications to oversight while drifting from intent. TM2 requires no capability that does not already exist.

Both models point to the same gap: the auxiliary deployment layer, where AI systems meet institutional pressures, legacy infrastructure, human operators, and real-world incentives, is where loss of control is most likely to emerge and where current monitoring has the least visibility. All four research workstreams in the paper independently converged on this finding.

The expert interviews are the empirical core

The interviews were not supplementary; they reshaped the analysis in several places. The inference chokepoint finding (that frontier reasoning is concentrated among a small number of cloud providers globally, making outbound traffic to inference provider endpoints the single most reliable detection strategy currently available) emerged directly from practitioner judgement. So did the convergence on output patterns, particularly over-thoroughness and capability-goal mismatch, as the most reliable behavioural discriminators. So did the structural finding that AI-modified code is increasingly not contributed back to public repositories, limiting the forward-looking signal value of public repository activity. So did the human over-trust finding: as one practitioner put it, "We will definitely get into interesting scenarios where AI just did things, and humans allowed them to do that because they got lazy."

Several practitioners came from social science and conflict analysis backgrounds rather than computer science. One observed: "It's not just a tech thing at all." This is consistent with our finding that effective monitoring requires people who understand social and institutional dynamics, not only those who understand the underlying models.

What needs to be built

The paper proposes a federated international monitoring capability, anchored in OSINT methods, designed to work with AI Safety Institutes as institutional bridges to closed data tiers, sustained by funding independent of the entities being monitored.

Limitations are real and stated throughout. OSINT cannot detect loss of control in real time or with high confidence. The most diagnostic signals sit behind access tiers that only regulatory mechanisms or structured data-sharing arrangements can open. The frameworks developed in the paper are exploratory and have not been validated against real-world incidents, because none have been confirmed.

These limitations do not argue against building the capability. They argue for building it now, on imperfect but available methods, while the detection window remains open. As deployment configurations move from cloud-hosted to hybrid to fully local, the external detection surface narrows. The OSINT window is not closed, but it is closing, which strengthens the case for building monitoring capability while these surfaces remain accessible.

OSINT itself did not arrive as a finished methodology. It was constructed by practitioners working outside formal intelligence institutions, in response to technological shifts that moved faster than any top-down framework could keep pace with. The discipline absorbed tradecraft from journalism, conflict monitoring, and forensic analysis as those sources became relevant, and it did so without waiting for permission. AI loss of control detection has the same structural feature. The technology is moving faster than any methodology built top-down can track, and the practitioners best positioned to build detection capacity are those already comfortable assembling methods in real time.

Implications

- **For OSINT and CTI practitioners**: the detection problem requires a methodological shift, not just new tools. Trace visibility depends entirely on how an agent is deployed and governed; assess deployment configuration before selecting collection methods. Treat agents as human actors for security purposes; searching for AI-specific technical signatures is a diminishing-returns exercise.

- **For policymakers**: fund independent AI monitoring capacity with ring-fenced, multi-year commitments from non-industry sources. Explore structured data-sharing arrangements for AI incident signals, modelled on financial services fraud signal consortia. Investigate whether existing data retention frameworks can accommodate AI-relevant traffic patterns as a lower-cost first step than new legislation.

- **For AI safety researchers**: the empirical base for detection work is missing. Design red-team exercises specifically for autonomous AI threats. Advocate for security operations centres to begin recording whether detected threats exhibit AI-associated behavioural signatures. Develop an AI-specific TTP taxonomy comparable to MITRE ATT&CK, including the proposed Capability Concealment tactic that has no current equivalent.

- **For funders**: independent technical monitoring capacity is the highest-leverage structural intervention identified by this research. The case for building now rather than waiting is that the detection window is open and the relevant capability thresholds are approaching. The alternative is not caution; it is arriving too late.