SPAR FALL 2026 · PROJECT PROPOSAL

Mitigating Intentional Loss of Control Risk Through Interoperability Standards for Agentic AI

Agent identity is the chokepoint that decides whether an autonomously deployed agent can be refused, revoked, and attributed. That layer is being specified now, in rooms with no loss-of-control threat model in them.

MENTOR	BASED IN	MENTEES	MENTOR TIME	AREA
Kevin Kohler	Geneva, Switzerland	1–2 · 10–20 hrs/week	2–3 hrs/week	International governance · AI security

1. The problem

Standards are where technical detail quietly becomes de facto international law. The Internet is the cautionary case: IP won the protocol wars on network effects, and its identity and trust layers were bolted on afterwards. Thirty years later BGP still runs on trust, the web PKI still depends on the good behavior of every certificate authority in the store, and the neutrality of the DNS root still turns in part on which government holds jurisdiction over ICANN. None of it has been fixable, because change control was settled long before anyone understood what was at stake.

Agent-to-agent interoperability is at that stage now, and moving faster. As of July 2026:

- **MCP**, the most widely adopted tool-use protocol, moved under a Linux Foundation body in late 2025. Its authorization model is conventional OAuth; the specification has no concept of agent identity.
- **A2A**, the leading agent-to-agent protocol, reached its first stable version under the Linux Foundation. An agent identifies itself through an Agent Card it publishes about itself, and signing that card is optional under the specification.
- **Agent payments** are moving fastest of all — Google's Agent Payments Protocol was donated to the FIDO Alliance in 2026. It is also the partial counterexample worth studying, because it anchors trust in cryptographically signed user mandates rather than in agent self-declaration.
- At the **IETF**, agent identity work remains a scatter of individual drafts, none adopted by a working group. A proposal to open a working group on agent delegation and traceability was declined in July 2026.
- **ITU-T** established a focus group on trust and identity for agentic AI only in July 2026, with first work items targeted for 2027 and beyond.
- **ISO/IEC** has no agent-specific standard in development.

The layer that will determine which agents the world's systems are willing to deal with is being written by a handful of individual drafts, one industry foundation, and a payments alliance. The window to shape it before network effects dominate is a few years at most.

2. The safety argument

Intentional loss of control is the case where a human deliberately deploys an agent to operate without an accountable controlling principal — designed to resist shutdown, to re-establish itself elsewhere, or to replicate. As capability diffuses through open weights, mitigation at the model layer weakens: you cannot rely on a developer's safeguards for a system whose weights are public and whose fine-tuning is cheap. What remains are the downstream chokepoints. In an economy where agents transact with agents, the chokepoint is **admission to the network**.

The architecture of agent identity decides whether that chokepoint exists at all:

- **Self-declared or self-attested identity.** The agent presents a credential it generates about itself. An adversarial deployer forges or clones it, re-registers after any block, and pays no cost. Revocation has nothing to bite on. There is no chokepoint.
- **Issued, cryptographically bound, revocable identity with an auditable delegation chain back to an accountable legal person.** Revocation becomes an intervention at the transaction layer rather than a request to a developer. Unattributed agents can be rate-limited or refused by counterparties, payment rails, and compute providers.

Liability attaches to someone. Incidents can be reconstructed. “Know your agent” becomes implementable rather than aspirational.

> The claim this project tests: organized, issued agent identity is a control lever; self-declared identity is not.

It is not a free lever, and the project has to be honest about that or it will not survive contact with the standards community. The same identifier that lets you revoke a replicating agent lets a state bind every agent action to a real-world human identity — the flexible-addressing problem of Internet architecture, one layer up. A registry is a governance object with an owner and a jurisdiction, which is the ICANN question again, with the same answers available (neutral host state, immunity from unilateral sanction, federated issuance) and the same failure modes. And a heavy identity requirement can serve market foreclosure as readily as safety. The question is therefore not whether agent identity is good, but:

> Which architecture buys the most loss-of-control mitigation per unit of surveillance capability, centralization, and foreclosure risk — and who must hold change control for that to remain true?

3. Research questions

1. **Change control.** Who holds decision rights over the identity and trust layer of agentic interoperability today — by body, document, and decision date — and where does that move over the next eighteen months?
2. **Requirements.** What does an intentional-loss-of-control threat model demand of an agent identity architecture, derived from concrete scenarios rather than from enterprise IAM best practice?
3. **Comparison.** How do the live design families score against those requirements, and against the counterpart criteria of privacy, centralization, jurisdictional exposure, and adoption feasibility?
4. **Layer assignment.** Which properties must live in the protocol, and which can be left to policy, procurement, or platform terms? Following Collingridge: what is cheap to change now and prohibitively expensive later?
5. **Intervention.** What is the minimum viable ask a safety-motivated actor could bring into a named process in the next twelve months, and to whom?

4. Work packages

WEEKS	PACKAGE	CONTENT
1–3	**WP1** Change-control map	Inventory the live artifacts and the bodies behind them, across the protocol foundations, the internet standards bodies, the payments alliances, the identity community, and the intergovernmental track. For each: who owns it, what stage it is at, how change control works, who funds and staffs it, and when the next decision falls.
3–6	**WP2** Threat model to requirements	Build three to four concrete intentional-loss-of-control scenarios — a deliberately released self-replicating agent operating across jurisdictions, an agent that survives shutdown by re-registering under a fresh identity, a delegation chain subdelegated until accountability is untraceable, a small-scale rogue internal deployment. Walk each through the transaction lifecycle from discovery to revocation, and derive the properties an identity architecture would need in order to interrupt it.
5–9	**WP3** Comparative assessment	Score the design families against those requirements and against the counterpart criteria — privacy, centralization, jurisdictional exposure, adoption feasibility. The families run from purely self-declared credentials, through certificate and verifiable-credential approaches with federated issuance, to registry-issued identifiers and payment-anchored mandates. Treat “who issues, who revokes, under whose jurisdiction” as a first-class variable rather than an implementation detail.
9–12	**WP4** Intervention and delivery	Map the findings onto WP1's decision calendar: which ask, which body, which deadline, which coalition. Produce one brief for standards participants and one for diplomats and AISI staff, who need the same

WEEKS	PACKAGE	CONTENT
		conclusion in a different language. Stretch goal, where a process is open: draft the comment or contribution text itself.

5. Deliverables

- A **change-control map** of the agent identity and trust layer, with a decision calendar. No public equivalent currently exists.
- A **requirements set** derived from an explicit intentional-loss-of-control threat model.
- A **scored comparative assessment** of the design families, with the dual-use trade-offs stated rather than buried.
- A **policy brief of roughly 3'000 words** for a Geneva and AISI audience, plus a shorter version for the technical standards community.
- Stretch: a **submitted comment or contribution** to a named open process.

6. Method

Document analysis of primary standards artifacts — drafts, charters, minutes, ballots, all public. Structured expert interviews, target eight to fifteen, across standards participants, identity engineers, and AI safety researchers, using the mentor's network for access. Scenario-based reasoning rather than quantitative modeling. Comparative scoring against published criteria with explicit uncertainty, so that a reader who disagrees with a weight can recompute rather than argue. No compute required.

7. Mentee profile

Strong fit: someone who can read protocol specifications and IETF or W3C drafts closely without needing to implement them — a policy or governance researcher with real technical literacy, a security or identity engineer curious about governance, or a lawyer comfortable with technical text. Interest in AI safety and catastrophic risk. Able to produce clear written output with limited supervision. Willing to email strangers and run interviews.

Not required: machine learning research experience, publications, prior standards participation, or a technical degree.

Team: one to two mentees at 10–20 hrs/week. With two, the natural split is WP1 and WP3 (landscape and comparison) against WP2 and WP4 (threat model and intervention), with a shared weekly synthesis.

8. Why this can land

First, I have worked on the politics of interoperability standards before. At the Center for Security Studies at ETH Zurich I published *One, Two, or Two Hundred Internets? The Politics of Future Internet Architectures* (2022), a study of change control, interoperability, and unintended political effects in Internet protocol design — New IP, SCION, the jurisdiction question at ICANN, the standards conflict between the IETF and the ITU. This project applies that study's method to agents, twenty years earlier in the technology's life.

Second, I am embedded in International Geneva, where I am launching a new AI governance institute this autumn. Together with an industry standards consortium I organized a private roundtable on AI standards that brought major AI labs, global standards bodies, a leading payment network, and three national AI safety institutes into the same room. It generated follow-up interest, and this project is designed to feed it. The findings go into those rooms rather than onto a preprint server — that is the theory of change.

Scoping note. I am not a standards engineer. The project is deliberately built around the political-economy, change-control, and threat-model questions, where that is not the binding constraint. Technical review is handled by putting drafts in front of identity engineers from the roundtable network before anything is published.