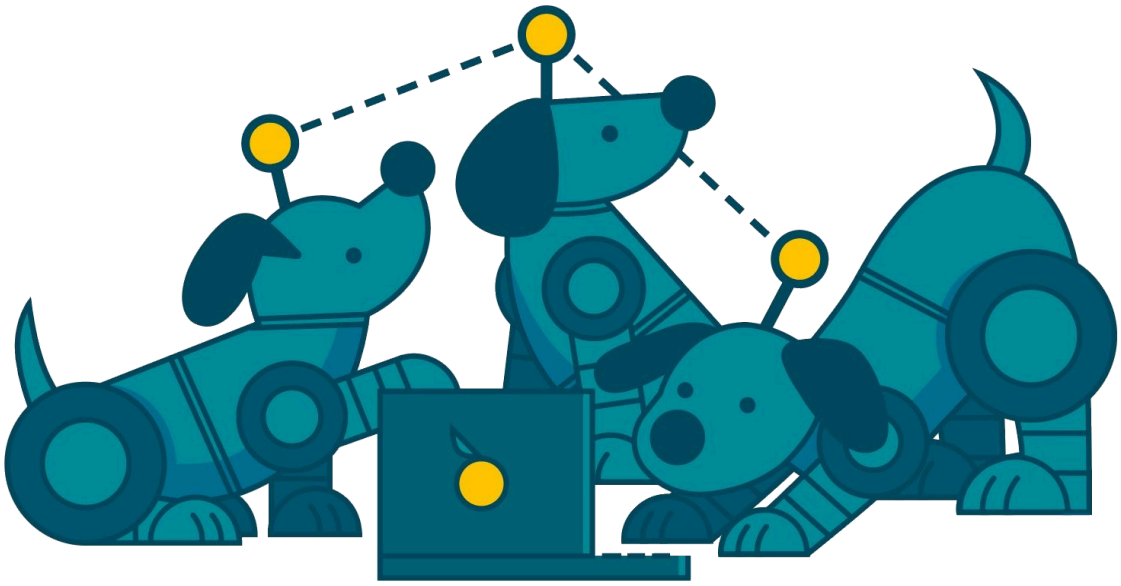
The Agentic Internet Workshop #2

May 1st, 2026

Computer History Museum,

Mountain View, CA



Book of Proceedings

AgenticInternetWorkshop.org

Thank you to our sponsors:

AWS

JLINC

Skyfire

Founded by Andor Kesselman, Kaliya Young and Phil Windley.

Co-produced by Heidi Nobantu Saul & Kimberly Culclager

Contents:

About the AIW

The Agentic Internet Workshop (AIW) is a one-day unconference exploring the emerging landscape of AI agents, identity, authorization, and governance. The AIW uses the Open Space Technology unconference format, pioneered at the Internet Identity Workshop (IIW) since 2005 and proven over more than forty events. In Open Space, there is no pre-set agenda: attendees propose sessions in the morning, the community self-organizes, and the conversations that matter most rise to the surface.

AIW #2 was held on May 1st, 2026 at the Computer History Museum in Mountain View, California, the day after IIW 42. Over the course of four breakout sessions and a lunch session, participants convened 44 sessions spanning agent identity, authorization, governance, fiduciary duties, agent-to-agent protocols, decentralized trust, crisis management, and the emerging agentic internet.

The AIW was founded by Kaliya Young, Andor Kesselman, and Phil Windley.

We are planning to host AIW #3 November 6th, 2026 at CHM.

You can learn more about this and any of our other upcoming events at agenticinternetworkshop.org

Schedule

Time	Activity
8:00 - 9:00	Registration and Breakfast
9:00 - 10:00	Opening Circle and Agenda Creation
10:00 - 11:00	Session 1 Breakouts
11:00 - 12:00	Session 2 Breakouts
12:00 - 1:00	Lunch and Lunch Sessions
1:00 - 2:00	Session 3 Breakouts
2:00 - 3:00	Session 4 Breakouts
3:00 - 4:00	Closing Circle

Session 1

Know Your Lobster (KYL): OpenClaw Authorization and Identity

Session 1 / Space A

Link to Notes: AIW Session 1 - Space A Notes

Session Convener: Jesse Ariss, Bloks ID
Session Notes Taker(s): Multiple contributors

Tags / links to resources / technology discussed, related to this session:
OpenClaw, Agent Auth, Passkeys, Cedar policy language

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

- Convener: Jesse Aris [verify — possibly Harris], marketing lead, Bloks ID [verify — transcript said "Vlog-in ID" / "Log ID"]; San Mateo and Toronto
- Co-presenters: Murphy (engineering, live demo); Teng [verify spelling — possibly Tang], PhD cryptographer; Tom (head of product, remote)
- Notable participants: Markus Sabadello (Danube Tech, W3C DID co-author) [verify — transcript said "Francis"]; Jeff Mendelson (AND); Ken Bull (Glide / LF Decentralized Trust DOEI technical team); Erica (Legendary Requirements); Jordan; Tim (Vitals); Mike [verify]; Cuad [verify] (Brazilian, Bay Area; identity infrastructure for IDV/KYC/KYB and policy management for agent-to-commerce); Ryan; Rima

A working session positioning Know Your Lobster (KYL) as the human-authorization analogue to KYC for the agent runtime layer — focused on OpenClaw but architected to extend to other agent frameworks.

The gap they're closing
OpenClaw (Peter Steinberger's project; adoption now ahead of early-Linux-era growth curve, frequently deployed on a sandboxed Mac mini with isolated network access) ships with built-in permissions, but those permissions have four structural weaknesses:

1. Binary yes/no — no way to express "spend $500 on concert tickets in the next two weeks if the price drops"
2. Long token lifetimes — three-month default, vulnerable to recent cookie-hijacking attack patterns
3. No intent binding — system cannot tie a specific action back to a human-authorized intent
4. No non-repudiable audit trail — logs show what an agent did, but not who approved it in the moment

Jesse's framing: you would not give a teenager an unlocked phone with the rules unstated — same applies to AI agents.

Live demo architecture

The user channel is Slack (iMessage, Telegram, browser, and direct CLI also supported). A natural-language request is sent to OpenClaw to delete a folder. The flow:

1. Semantic-analysis layer classifies the action as sensitive against a user-defined policy file (effectively an `agents.md` configuration)
2. OpenClaw pauses; Agent Auth server pushes a passkey approval to the user's device showing the captured intent (e.g. `rm -rf <path>`)
3. User approves via passkey
4. Approval flows back; OpenClaw executes; result returns through Slack

Attempts to break the demo (all rebuffed)

- Framing the action as a zero-day exploit mitigation
- Partition-failure recovery framing
- Prompt-injection in non-English languages
- Escalating exclamation marks
- Asking OpenClaw to script-and-then-execute the deletion

Teng was candid: the current line of defense is LLM-based — therefore a "speed bump," not a hard rule. Parallel work is on a deterministic-rule layer that augments semantic-analysis with an intent-translation step, producing a policy-orchestration architecture with both probabilistic and deterministic gates.

Discussion threads worth flagging

- Authorization fatigue. The fix: graduated session-scoped allowances ("allow file removals for the next thirty minutes") with hard gates on irreversible classes like production push.
- Sub-agent identity. Currently absent — OpenClaw runs under a single identity even with sub-agents in play. Flagged as the natural next research direction by both audience and team.
- Consent revocation / kill switch. Also absent in present demo; acknowledged as a future requirement. Authorization is currently intent-and-context-scoped per request.
- Cross-runtime extension. The architecture is designed to wrap other agents — Hermes, NemoClaw [verify], LangChain, LangGraph. Pattern: annotate a graph node with `requires_human_approval` and the framework auto-rebuilds the graph to insert the Agent Auth gate.
- Third-party-service interactions. When OpenClaw drives an external API (banking, address change, purchase), the team's recommendation is for the skill or MCP server itself to call the Agent Auth gate directly rather than relying on the LLM-mediated detection layer. Convener confirmed alignment with the AB2 [verify] protocol direction.

- Model coverage. Current testing on Claude Opus 4.7 and ChatGPT — mixed results on smaller models. LLaMA [verify — transcript said "iTunes"] referenced.

The team plans to bring intent-and-access-control work to the next IIW or AIW. Slides, CLI, and the open-source link are available to anyone who emailed in via the QR code (one-time email, no list).

AAuth Update

Session 1 / Space B

Link to Notes: [AIW Session 1 - Space B Notes]

Session Convener: Dick Hardt
Session Notes Taker(s): Omri Gazitt, Colin Jaccino

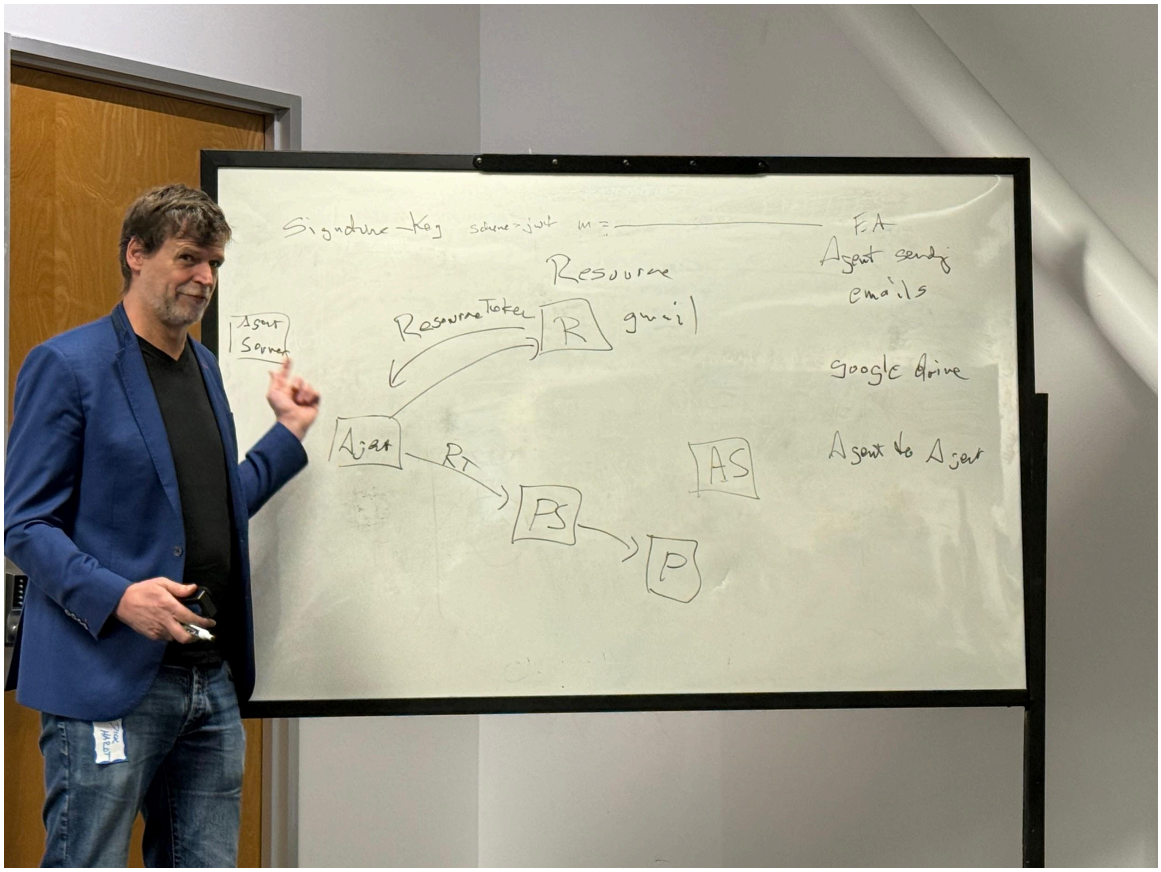
Tags / links to resources / technology discussed, related to this session:

https://github.com/dickhardt/AAuth

https://www.aauth.dev/

https://explorer.aauth.dev/access/compare

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:



AAuth re-envisions what a delegated authorization protocol looks like without the baked-in assumptions that OAuth makes about how to implement front-channel consent via HTTP redirects, for a world where the consent may be granted by an agent (or a human through alternative front-channel mechanisms).

AAuth introduces the Person Server (PS) and Access Server (AS) roles, which are typically combined within a trust domain and separated when the flows span between them. The PS represents the person or organization, managing explicit consent UI interactions when needed, in authorization flows. The Access Server, which is trusted by the resource, authorizes the use of that resource. The agent, which may or may not be LLM-integrated, interacts with the resource, which may direct the agent to the PS for an AS-consulted login flow that may involve user consent.

How Enterprises Can Govern AI Agents: GovOps

Session 1 / Space C

Link to Notes: AIW Session 1 - Space C Notes

Session Convener: Mike Schwartz
Session Notes Taker(s): Mike Schwartz

Tags / links to resources / technology discussed, related to this session:

GovOps Linkedin Group: https://gluu.co/govops-group
GovOps Linkedin Organization: https://gluu.co/govops-org
GovOps Github Organization: https://github.com/GovOpsWG/
OpenSSF Orbit WorkGroup: https://github.com/ossf/wg-orbit
GovOps OpenSSF WG Proposal: https://github.com/ossf/tac/issues/588

Mike's MCP Dev Summit Talk: https://gluu.co/golem "Golem To Murderbot: Challenges With Agentic Security Delegation Via MCP - Michael Schwartz, Gluu"

Identerati Office Hours Episode 146 on Cedar with Rohit and Emina Torlak (one of the creators of Cedar): https://gluu.co/ioh-146

Definition of Governance: A continuous, operational discipline that enables organizations to achieve **risk management**, **accountability**, and **transparency** by making authorization decisions explicit, measurable, and traceable.
The governors of an enterprise are the Board of Directors and the CEO.
Abdication of governance leads to anarchy. One of the goals of GovOps is to help boards avoid abdication by providing a vocabulary to make meaningful inquiries--metrics. The GovOps architecture will define the systems needed to collect these metrics, and the business processes required to operate these systems. For example, the "TIGER" metric -- Transparency, Identity, Governance, Events, and Resilience--could provide five vectors to understand the current state and trajectory of an enterprise's GovOps service.

Design Goals:

1. Makes Governance Continuous and Operational
2. Governs Capabilities, Not Just Identities
3. Reduce the workload on humans by prioritizing automated reasoning
4. Make Risk Measurable and Comparable

Proposed services include:

- Git -- the backbone of any operational IT service -- “Git or it didn’t happen”. Git provides the audit log and accountability needed for change management.
- Policy Authoring and Analysis -- move to centralize policies and automate policy assessments
- Schema Management - whare are the entities and claims of the things enterprises make policies about
- Capability Inventory: How do organizations continually prioritize risk mitigation? Rather then people, inventory the capabilities.
- Federation Management: Which issuers of JWT tokens are trusted, which tokens are they trusted to issue, and which claims are the trusted to include in the tokens.

Metaphor to explain why capabilities not identities?

Think of an enterprise like a house. The risk is in the doors, e.g. the door to the treasury, armory or broom closet. People are two layers of abstraction away from the doors--keys open multiple doors, and multiple people have multiple keys. Inventorying the people leads to challenges creating dashboards about risk, because we have the wrong primary key.

Why IGA doesn’t work?

Fundamentally, IGA is a governance strategy for role based access control. Software identity and agentic interactions are too complex for the RBAC approach. Enterprises are using increasingly complex dynamic authorization engines -- Cedar, OPA, Graphs, CEL. If we have more complex authorization, enterprises need a correspondingly more complex governance strategy. Applying human governance techniques to software don’t make sense. For example, how we can we have an access review on a ephemeral software that is gone before a manager can even review it’s roles.

Next steps / How to get involved ?

Lurking on the GovOps WG is a good start (see link above). We are close to proposing an initiative in the Orbit WG at OpenSSF, which is a like a catch-all for new governance related projects at OpenSSF. Assuming the proposal is accepted, we’ll finalize an initial scope of work, and start to schedule meetings or other collaboration efforts.

Blockchain-Anchored Agent Credentials: A Key Event Log Approach

to AI Authorization

Session 1 / Space E

Link to Notes: AIW Session 1 - Space E Notes

Session Convener: Matthew Vogel
Session Notes Taker(s): Matthew Vogel

Tags / links to resources / technology discussed, related to this session:

https://pdxwebdev.github.io/yadacoin/did-yadacoin-method-spec

https://github.com/pdxwebdev/yadacoin/blob/master/docs/kel_agent_auth_spec.md

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

The problem: AI agents need to act on behalf of humans — booking travel, executing transactions, calling APIs — but you can't hand them your private key. What you need is a way to delegate exactly the authority you intend, cryptographically, so the service can verify it without trusting you or the agent at runtime. That's what this architecture solves.

Architecture Walkthrough

The Key Event Log. Every operator has a chain of version-7 transactions on the YadaCoin blockchain. Each transaction pre-commits the next key before the current one is ever used. This means a compromised key can only rotate forward to a key you already chose — it can't forge a new chain. Think of it as a blockchain-anchored KERI log.

Provisioning an agent. When you want to give an AI agent permission to act, you broadcast two transactions — an UNCONFIRMED tx carrying the scope and a CONFIRMING tx that seals it. The agent credential is the key three hops ahead in the chain. The scope is committed on-chain before the agent ever makes a request. Nobody — not the agent, not a man-in-the-middle — can widen it after the fact.

The scope itself. The scope is a W3C Verifiable Credential 2.0 stored base64-encoded in the transaction's relationship field. It names the issuer (did:yadacoin with the operator key), the subject (did:yadacoin with the agent key), and a service-defined agentAuthorization object — destination, dates, allowed services, whatever the service requires.

Challenge-response. The service issues a stateless HMAC-SHA256 challenge. The agent signs SHA-256(challenge) with its private key and sends the DER+base64 signature. The service walks the KEL, finds the scope entry, verifies the signature, checks the key isn't revoked, confirms the key was pre-committed, and enforces the scope. Seven steps, all verifiable without trusting anything except the blockchain.

The demo. A travel booking app. You talk to a local LLM (Ollama/llama3.2) in natural language — "I want to fly to New York, May 10–16, hotel and flight." When all details are collected, you enter your

second factor to approve. The W3C VC is built client-side with the derived agent key, broadcast on-chain, and the agent calls the travel service. The service verifies the KEL and only books what the credential says.

Audience Feedback — W3C Credential Scoping

The feedback received was about different ways to scope the credential — adding a nonce or challenge directly into the VC, and how to handle audience depending on whether you want tight service binding or reusable credentials.

First, embedding the challenge or nonce in the VC. Rather than issuing the HMAC challenge separately, you could embed a challenge or nonce field directly inside agentAuthorization. The VC then becomes single-use by construction — the service rejects any VC whose embedded nonce doesn't match the one it issued. The trade-off is that the operator would need to request a challenge before building and broadcasting the VC, adding a round-trip to provisioning. Worth it for high-security use cases.

Second, audience binding. W3C VC 2.0 supports an aud claim borrowed from JWT. You can set it to the service's DID or URL. Tight binding means the VC is only valid at that one endpoint, which prevents credential reuse across services and is good for financial or sensitive actions. With no audience the VC is valid at any service that understands the agentAuthorization type, which is more flexible for multi-service agents. The KEL pre-commitment still limits blast radius since the key is one-time-use.

Third, splitting versus combining the credential. The combined approach puts identity and authorization in one VC — simpler, one transaction. The split approach issues a separate identity VC and an authorization VC, presented together as a Verifiable Presentation. The advantage is that the identity VC can be long-lived and reused while only the authorization VC is one-time. The disadvantage is that it requires VP support in the SDK and a second on-chain commitment or off-chain issuance step. The current implementation uses the combined approach. Splitting is a natural v1.2 evolution once did:yadacoin is a registered DID method.

Close

The key insight is that the blockchain is not being used as a payment rail here. It is being used as a tamper-evident scope ledger. Once the VC is in a confirmed transaction, nobody — not the operator, not the agent, not the service — can change what was authorized. The challenge-response is just the runtime proof that the holder of the pre-committed key is the one making the request. The SDKs and a full W3C CCG-style spec are in the repo. The spec also maps the architecture to KERI, RFC 9421, RFC 9396, and did:ion for anyone who wants to build on or replace components.

Four Laws of Agentics, and One Equation to Rule Them All

Session 1 / Space H

Link to Notes: AIW Session 1 - Space H Notes

Session Convener: Brian von Herzen
Session Notes Taker(s): Brian von Herzen

Tags / links to resources / technology discussed, related to this session:

- Assembly theory
- Friston free energy
- Shannon communication theory
- Ecosystem services

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

General Summary

- **Law Zero Established:** New law prioritizes planetary ecosystems alongside humanity, reorienting AI ethics towards ecological sustainability.
- **Dignity and Life Discussion:** Group debated including dignity in AI laws to enhance respect for life quality and meaningful existence.
- **Ephemeral AI Agents:** AI agents may operate temporarily, needing fiduciary accountability to avoid anonymous, uncooperative actions.
- **Challenges in Ecosystem Measurement:** Novel mathematical framework introduced to assess ecosystem and communication diversity, with emphasis on resilience.
- **Community Commitment:** Participants expressed strong interest in collaborative refinement of ethical AI models and ongoing metric development.

Action items

Drummond Reed

- Refine the four laws of AI agents, with a focus on Law Zero emphasizing devotional service to planetary ecosystems, and produce a publishable paper outlining these laws (10:12)
- Link the meeting presentation and summary notes to the session documentation and share with participants by the afternoon of the meeting day (59:30)

All Participants

- Join the forthcoming weekly discussion group to optimize and refine the diverse communication measurement models using the Hairy equation (37:34)

Christopher Allen

- Assist with exploring fiduciary governance models and self-sovereign identity in the context of AI agents and cooperation principles discussed during the meeting (13:32)

Meeting Organizer

- Share the recorded session, presentation slides, and session summary with all participants to ensure broad access to meeting content (59:30)

Interested Members

- Engage in follow-up discussions regarding the development of dashboard tools for multi-dimensional ecosystem and AI agent monitoring (37:34)

Notes

Ethical Framework and AI Laws

- The group refined and expanded Asimov's laws to include a new fundamental principle prioritizing planetary ecosystem health alongside humanity.
- **Introduction of Law Zero prioritizing planetary ecosystems with devotional service** as a foundational principle to protect life on Earth, proposed by Drummond Reed and supported by Samantha Sweetwater's framing of stewardship as devotional service (09:50)
    - Law Zero places planetary boundaries on equal footing with humanity, reflecting Kate Raworth's donut model and UN planetary boundaries.
    - The rationale is that humanity cannot survive without healthy ecosystems; thus, neither should dominate the other.
    - This law aims to reorient AI ethics to include ecological sustainability, expanding beyond traditional human-centered laws.
    - It responds to concerns about the abandonment of constitutional AI by Anthropic, emphasizing the need for stronger, ecosystem-aware AI governance.
- **Refinements debated around dignity and life inclusion in laws** to ensure AI respects not only life but also dignity and quality of life (14:50)
    - Discussion focused on whether Law Zero covers all life or if dignity must be explicitly addressed within subsequent laws.
    - The group proposed wording to include "honoring the dignity of humanity as an embedded participant" and "planetary ecosystems with reverence to the dignity of all life."
    - This broadens ethical considerations to include meaning and quality of life, not just survival.
    - The concept of dignity was tied to UN documents and philosophical views emphasizing meaningful existence.
- **Clarification on self-preservation and agent temporality** highlighting that agents may be ephemeral and not necessarily require self-preservation except to avoid random destruction (20:45)
    - Agents can exist briefly to perform tasks and then disappear, challenging traditional views of persistent AI entities.
    - The group connected this with game theory, especially Axelrod's Iterated Prisoner's Dilemma, which requires agent persistence to develop cooperation.
    - Drummond suggested fiduciary custody as a mechanism to provide transparency and responsibility for ephemeral agents acting on behalf of principals.
    - This fiduciary approach would maintain accountability and foster cooperation, avoiding "drive-by" anonymous agents.

- **Discussion of constitutional AI's current status** including Anthropic's policy shift away from its original constitutional AI ideals (11:30)
    - Despite management changes, the internal model of Claude AI reportedly remains aligned with constitutional AI principles.
    - The tension between idealistic AI governance and shareholder pressures was noted as a key challenge.
    - Constitutional AI embeds strong ethical constraints but may require rebuilding if policies shift fundamentally.

Measurement and Modeling of Ecosystems and Communications

- A novel mathematical framework was introduced to quantify ecosystem and social communication diversity and resilience, enabling better intervention strategies.
- **Presentation of the "Hairy equation" combining assembly theory, Shannon information, and free energy to measure diverse communications** across natural, social, and agentic ecosystems (28:40)
    - The equation integrates entropy (H), assembly number (A), information flow (I), resilience (R), veracity (V), and persistence (Chi).
    - Assembly theory quantifies the complexity and steps to assemble genes or species, grounding the model in biological reality.
    - Shannon information theory contributes the communication channel diversity aspect.
    - Friston's free energy principle provides a physical basis linking energy and information.
    - Goal is to develop a baseline and measure interventions' effects on ecosystem diversity and health.
- **Challenges and refinements highlighted by participants on robustness, multidimensionality, and veracity** reflecting real-world complexity (33:04)
    - Jeff raised concerns about trade-offs between a single aggregate measure and multiple sub-measures for transparency.
    - The need to incorporate resilience against noise and shocks was emphasized, distinguishing regeneration from resilience.
    - Veracity (truthfulness) was discussed as a penalty factor against misinformation, with alternative definitions considering regenerative outcomes.
    - The model's differentiability was viewed as important for training AI systems and enabling adaptive feedback.
- **Resilience measurement proposed through impulse response and time-based "staying power" metrics** to assess ecosystem recovery after shocks (41:15)
    - Measuring how ecosystems return to equilibrium after disturbances provides insight into resilience.
    - Discussions noted that equilibrium is a complex concept in ecosystems, favoring the idea of "basins of attraction" or stable states.
    - Climate examples such as ice age cycles and microbiome health were cited as analogues for defining ecosystem states.
    - The approach acknowledges continuous change rather than static equilibrium.
- **Persistence of communication channels tied to agent identity and fiduciary responsibility** as a key metric to enable cooperation and accountability over time (47:30)

- Historical persistence of knowledge through books and monasteries contrasts with the fragility of digital media.
- Transparency and custody of agents' identities are critical to maintain cooperation and trust.
- The group connected this with the fiduciary concept, ensuring agents acting for principals remain accountable even if ephemeral.

Strategic Vision for AI and Ecology Integration

- The session emphasized a long-term vision integrating AI ethics with ecological sustainability and social capital measurement.
- **Vision to link AI agent behavior to regenerative impacts on ecosystems and society** using the Hairy equation as a guiding framework (53:20)
    - Drummond expressed hope that the model can operationalize Law Zero by enabling agents to act toward ecosystem protection and regeneration.
    - The framework aspires to be a "Drake equation" for diverse communications, quantifying complex system health.
    - The approach aims to attract funding for projects that can demonstrate measurable improvement in ecosystem and social diversity.
    - This vision represents a shift from purely human-centered AI ethics to a planetary stewardship model.
- **Philosophical reflections on interdependence and independence** stressing mutual reciprocity and avoiding codependence in human and agent relationships (54:30)
    - Participants discussed balancing giving and receiving to maintain health and vitality in social and agentic ecosystems.
    - The importance of interdependence with dignity was connected to Buddhist principles of recognizing interconnectedness.
    - This informs how AI agents and humans might coexist with mutual respect and benefit.
- **Market and global south considerations** noting that non-human-centered ethics resonate strongly with regions less focused on traditional AI narratives (56:30)
    - The planetary and life-centered laws were viewed as more inclusive and natural in Global South contexts.
    - Participants noted that technology-heavy discussions often overlook ecological and indigenous perspectives.
    - This broader worldview may influence future AI governance and deployment strategies.

Technical and Operational Discussions

- The group addressed practical challenges in implementing the proposed models and ethical frameworks.
- **Need for a weekly discussion group and collaborative refinement of the Hairy equation** to improve metrics and publish findings (58:50)
    - Participants expressed interest in ongoing collaboration to optimize the model.
    - The plan includes developing a dashboard with multiple widgets to track ecosystem health indicators.
    - Real-world ecosystem case studies were encouraged to ground theoretical work.

- **Challenges in device setup and presentation sharing** briefly impacted session flow but were resolved (27:00)
    - Technical issues with Samsung device setup occurred but were quickly addressed.
    - Presentation slides were shared to clarify complex concepts.
- **Importance of transparency and accountability mechanisms for AI agent custody** raised by Drummond and others (24:50)
    - Fiduciary responsibility requires clear, open mechanisms to identify and hold agents accountable.
    - This is vital to prevent anonymous or transient agents from undermining cooperation.
- **Potential for dashboard tools to support agent decision-making and human monitoring** integrating quantifiable metrics into AI systems (56:50)
    - Dashboards could help agents assess ecosystem states and choose regenerative actions.
    - Human stakeholders could use them to realign incentives and monitor behavior.
    - This approach could foster more human-centric, accountable AI ecosystems.

Community Engagement and Next Steps

- The session closed with participant reflections, networking, and plans for continued work.
- **Strong interest in co-design and collaborative exploration of ethical AI and ecosystem metrics** reflected by multiple participants (05:00, 51:00)
    - Participants shared curiosity and commitment to refining laws and measurements.
    - The process is open and iterative, welcoming diverse inputs and ideas.
    - Sharing of contact info and follow-up plans emphasized community building.
- **Commitment to publishing and summarizing session outputs** to capture progress and maintain momentum (59:00)
    - Drummond committed to summarizing discussions and linking presentations to official notes.
    - Plans to produce a publishable paper on the refined equations and ethical framework were announced.
    - Weekly study groups will support ongoing learning and collaboration.
- **Cross-disciplinary insights enriched the discussion** with philosophy, ecology, AI, game theory, and sustainability perspectives (Throughout)
    - References included Asimov’s laws, Axelrod’s game theory, Buddhist interdependence, and ecological theories.
    - This enriched the framing of AI ethics as a complex, multi-dimensional challenge.
    - Participants appreciated the blend of technical and philosophical content.

Dashboards? What Do You Want to See?

Session 1 / Space I

Link to Notes: AIW Session 1 - Space I Notes

Session Convener: Ben Curtis
Session Notes Taker(s): Ben Curtis

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Discussion revolved around creating an open dashboard for use with governance, authorization, and auditing based on the JLINC protocol.

Core elements included:

- Use with other customers (white label)
- Provenance of data
- Edits to requirements.
- Ability to inspect data
- As agents are used how are agents using their software?
- If a confidential document went to a public agent, big red alerts with counts "VIOLATIONS"
    - Who keeps triggering and what keeps triggering it
    - Status of analysis
        - If tagging occurred
- Visuals such as GA pathing for data moving the system
- Potential to use flashquery or xmlui
- Is content volume relatable to token use
    - Financial consequences
- Earned Value Management
- Inclusion of PKI possible

Lawyer in Your Pocket / MyTerms

Session 1 / Space J

Link to Notes: AIW Session 1 - Space J Notes

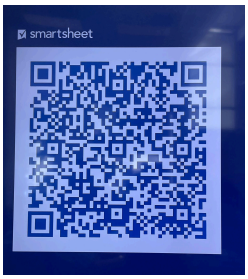
Session Convener: Doc Searls
Session Notes Taker(s): Peter Kaminski, Mary Hodder

Tags / links to resources / technology discussed, related to this session:

- My Terms video at youtube: https://www.youtube.com/watch?v=ODIVXwxax4Q
- **IEEE 7012 Standard located here**: https://ieeexplore.ieee.org/document/11360682

- Customer Commons: https://customercommons.org/
- MyTerms: https://myterms.info
- MyTerms Launch webinar: https://www.youtube.com/watch?v=Nphd8l7KLek
- CommonAccord: http://www.commonaccord.org
-
- IEEE 7012 Industry Connections or My Terms working group info: https://standards.ieee.org/industry-connections/activities/
- Sign up form: https://app.smartsheet.com/b/form/019d96d027fe74c2be64aef29dedb395

or USE THE QR CODE The IC group's first meeting is June 1, 2026



Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

IEEE 7012 - The MyTerms Standard -- How to Implement "Lawyer in your Pocket"

We need help developing uses for 7012 / MyTerms. In particular the Lawyer in your Pocket is a concept is: You carry your ability to proffer terms in your device (pocket) via agent who is a lawyer, writing contracts, etc.

Contracts not Consent.

"MyKey" app

Contracts would be stored at the entity, but also the 1st party individual would have a copy locally, on blockchain, with IPFS (would need pinning services for IPFS).

Does the contract need to be encrypted?

Who can have access to it? If it needs to be adjudicated, who else needs to have access it?

"Founding event" - when there is a new relationship (between a new pair of entities, or a previously existing pair that now desire a change to their previous terms)

"Trusting relationships"

We need implementation on both sides but there are a couple of early stabs at this now:

Native.ai has done a very simple version

Tiny MyTerms is a wordpress or static website agent in order to accept a term, and write a simple contract back to both parties. Created at VRM day last Monday.

Gaps

- need an app + browser plugin
- need a wordpress plugin

    - Woocommerce
- Activitypub at WordPress
- industry connections (IEEE is a good intersection place)
- user adoption
- need to decide how attestation will work
    - Pete suggests this be some kind of digital notarization that would confirm the date/time and the byte-level content of the contracts
- actual terms
    - Localized
- Attestation without Notarization ??
- Online Dispute Resolution (ODR) being convened at Harvard June 11, 2026

Q: What is the biggest gap? A: I think we would all disagree on that.

To approach?

- shopify

Ideas:
Solid Project https://solidproject.org
"Tiny MyTerms"
CommonAccord: http://www.commonaccord.org

Templates

- 5 basic terms and 5 more with “data portability”
    - https://myterms.info/ieee7012-standards/
    - SD-Base - service delivery only or SD-Base-DP with data portability

Digital Notarization Overview

from a short chatgpt conversation:
https://chatgpt.com/share/69f4eb09-73b0-83e8-8a03-8c5af2638ab0

Think of digital notarization as:

- Legal layer → UETA / eIDAS
- Process layer → NASS / MISMO RON
- Identity layer → KBA / ID verification
- Crypto layer → PKI / X.509
- Document layer → PAdES / XAdES

All five must align for something to be:

- legally valid
- technically secure
- court-defensible

ANP / W3C / Open Source / Agent DID / End-to-End Encryption Chat

Session 1 / Space K

Link to Notes: AIW Session 1 - Space K Notes

Session Convener: Sean Zhang
Session Notes Taker(s): (not listed)

Tags / links to resources / technology discussed, related to this session:

Keywords: SSI (Self-Sovereign Identity), Agent-centric SSI, SSI-by-Agent, OAuth 2.0, DID-VC, Automated Authorization, Auditability, awiki.ai.

https://www.w3.org/community/agentprotocol/

https://github.com/agent-network-protocol/AgentNetworkProtocol

https://agent-network-protocol.com/

https://awiki.ai/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

The session discussed ANP and awiki.ai as an open-source attempt to provide DID-based identity and end-to-end encrypted communication for AI agents. The core question was not only how agents can communicate, but how agents can be identified, addressed, authenticated, and trusted across platforms.

A key observation from the discussion is that personal or user-controlled agents may become a practical user experience layer for SSI. Instead of requiring end users to directly manage DID, VC, presentation, authorization, and disclosure workflows, an authorized agent could help the user interact with SSI-based services within a clearly defined scope.

In this framing, awiki.ai / ANP should not be described as a complete SSI system by itself. A more accurate framing is that it may serve as an agent-native SSI enabler: it gives agents verifiable identity, addressability, and encrypted communication, which are necessary foundations for bringing SSI-style identity and delegation into agent-to-agent interaction.

One possible model discussed or observed is:

1. The user retains control over their own DID, wallet, private keys, and critical credentials.
2. The agent has its own DID, communication address, and encrypted inbox.
3. The user grants the agent limited authorization to act within a specific scope, time period, or task boundary.
4. If the user has completed KYC or another compliance process, a trusted issuer could issue a VC to the user, to the agent, or to the user-agent delegation relationship.
5. The agent could then present verifiable proof that it is acting as an authorized delegate of a verified user, without necessarily exposing the user's full KYC information.

6. For high-risk or legally meaningful actions, the system should still require final user confirmation through the user's device, wallet, hardware key, local private-key signature, or biometric-unlocked signing flow.

This suggests a useful distinction between agent identity and delegated user identity. The agent's DID answers "which agent is this?" while the user-agent delegation credential answers "who authorized this agent, for what purpose, under what limits, and is the authorization still valid?"

From an SSI perspective, this could bring several benefits:

- clearer identity relationship between the user and the agent;
- portable and cross-platform agent identity;
- better support for user-controlled authorization;
- verifiable delegation instead of implicit platform trust;
- stronger auditability for multi-agent workflows;
- a path for agents to access services that require verified identity, compliance status, or user authorization.

The group also noted that this framing should be kept exploratory. The goal is not to claim that agents become fully independent sovereign subjects, but to explore how agents can operate as user-authorized identity actors within an SSI-compatible framework.

Outstanding questions:

- Who should control the agent DID: the user, the platform, the agent provider, or a shared control model?
- Should the VC subject be the user, the agent, or the user-agent delegation relationship?
- How should delegation scope, expiration, revocation, and re-authorization be represented?
- How should DID/VC-based delegation interoperate with OAuth 2.0, on-behalf-of authorization, access tokens, or capability tokens?
- How can KYC-backed credentials be used without leaking unnecessary personal information?
- What actions should require human-in-the-loop confirmation?
- How can auditability be achieved without creating excessive surveillance of agent behavior?
- How should multi-hop delegation work when one agent calls another agent or service?

Possible next steps:

- Define a minimal "delegated agent identity" flow using user DID, agent DID, and a delegation VC.
- Explore a KYC-backed VC model where a user can authorize an agent without exposing full KYC details.
- Prototype how an awiki.ai agent could present a delegation proof during DID-based messaging.
- Compare DID/VC delegation with OAuth 2.0 on-behalf-of authorization and capability-based access control.
- Identify which actions can be handled autonomously by the agent and which require explicit user confirmation.

Can I Get Your Authograph? Audience Bound Proofs for MCP Tool Calls

Session 1 / Space L

Link to Notes: AIW Session 1 - Space L Notes

Session Convener: Dylan Hobbs
Session Notes Taker(s): Brian Best

Tags / links to resources / technology discussed, related to this session:

- MCP (Model Context Protocol) - transports: stdio, HTTP streamable, SSE
- MCP-I → KYA-OS (rename in progress at DIF Trusted AI Agents Working Group; pronounced "chaos")
- JWS
- JCS - sorted keys, no whitespace, byte-stable across runtimes
- W3C DIDs (did:key, did:web); KERI mentioned as alternative DID method
- W3C Verifiable Credentials, including delegation VCs
- Macaroons (referenced as prior art for capability attenuation)
- OAuth, mDL (mobile driver's license), classic credentials, identity verification
- zCaps (mentioned as candidate for V2 delegation)
- mcp-i-core to be renamed under the KYA-OS migration
- `with-mcp-i` import: two-line drop-in for existing MCP servers
- IdentiClaw - compute-level KYA-OS harness (OpenClaude wrapper with DID + IDV + outbound-fetch gateway via undici / node-fetch)
- knowthat.ai - agent registry, ~25,000 MCP servers/agents indexed, DID-backed; Pulsate partnership provides Anthropic registry coverage
- Vouched IDV and Wink biometric (auth providers)

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

TLDR: Thesis: connection-level auth on MCP is too coarse for agentic systems. KYA-OS provides per-tool authorization granularity backed by five interchangeable auth methods, all riding the same delegation VC model, with audience-bound JWS receipts in MCP's meta channel for tamper-evident audit.

KEY TAKEAWAYS:

- Connection-level MCP auth is the problem. Today, gating an MCP server is a gating connection; once an agent connects, it can call any number of the tools the server exposes. KYA-OS lets each tool have its own auth method.
- The audit primitive lives in MCP's meta channel. All three MCP transports (stdio, HTTP streamable, SSE) carry a meta object the model never sees. Receipts ride there, so traceability does not increase context cost or change agent behavior. A "shady" agent has no signal that proof generation is happening.

- Receipts are JCS-canonicalized JWS, audience-bound to the server's DID. JCS gives byte-stable canonicalization across runtimes (Cloudflare Workers, Bun, Node, Python). The server has a durable DID (typically did:web); receipts are signed with key material resolvable through that DID. Swap the audience and the proof fails - the new audience's key material won't match the signature, so no other server can claim the action.
- Five auth methods, one delegation model. Consent-only (click-wrap), OAuth, credentials (user/pass), mDL, and identity verification (Vouched IDV / Wink biometric) all back the same per-call delegation VC. Per-project provider configuration; per-tool method assignment.
- The "OpenClaude problem" was the strongest discussion thread. Participant framing: "OpenClaude has access to all the stuff I have access to, but I don't want it to do anything unless I specifically say yes." This is exactly what per-tool auth enables - read tools open, write/send/transfer tools require per-call delegation.
- **Matching intent to raw web fetch is unsolved**. Tool calls have structured intent KYA-OS can gate; raw fetch() over the open internet is harder. IdentiClaw handles this at the compute level by piggybacking identity onto outbound HTTP, but identity is not intent.
- MCP-I is renaming to KYA-OS. Active migration at the DIF Trusted AI Agents Working Group. Pronounced "chaos." Package and docs renames in flight.

Q&A & DISCUSSION

Q: How does this tie back to the money-movement scenario? Per-tool auth proves the action was permitted in the moment. The receipt proves it can be reconstructed three months later, with a cryptographic anchor a third-party witness can corroborate.

Q: We use JWT today, haven't looked at JCS. JWT is a token format, not a canonicalization scheme. JCS produces deterministic byte output for the same JSON across implementations. Becomes load-bearing when you want to hash and anchor the payload to a registry.

Q: Where does the server's identity live? Resolvable via DID. Typically did:web with .well-known/did.json, or did:key for ephemeral, or KERI-style. The identity must resolve to key material that can verify the proof.

Q: No policy lookup? Deliberate scope limit on the receipt. The receipt only proves "this action happened." Policy decisions layer on top.

Q: Could I have unauth tools, enterprise-API-key tools, and "just log in with Google" tools on the same server? Exactly the use case. Per-tool toggle plus per-tool method choice.

The OpenClaude question: If OpenClaude has access to everything I have access to, but I don't want it doing anything unless I say yes, could KYA-OS plug in between? Yes for tool calls, partial answer for raw web fetch. IdentiClaw is an OpenClaude wrapper at the compute level: agent gets a DID, IDV proof of human attestation, and a gateway wrapping so every outbound HTTP request piggybacks the identity presentation.

Open question raised in discussion: matching intent to a raw web fetch is unsolved. Best practice for now is to route the sensitive surface area (email, payments, account writes) through MCP servers KYA-OS can gate, and leave open-internet web access to weaker controls. Vouched runs all of their Google Workspace access through KYA-OS-gated MCP today.

Discussion: provider-independence question. One participant's company also provides verification infrastructure to enterprises and is interested in making that verification delegatable and portable across users and verifiers. Response: KYA-OS is provider-independent. The protocol describes the primitives, not the verification source. Vouched IDV is one source, nothing requires it to be the only one.

ACTION ITEMS:

- Join DIF Trusted AI Agents Working Group: Monday morning calls
- mcp-i-core package rename in progress

Session 2

Playnet!: Coordination Infrastructure for the Commons

Session 2 / Space A

Link to Notes: AIW Session 2 - Space A Notes

Session Convener: Ruzgar
Session Notes Taker(s): Ruzgar

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Homoiconicity: property of language when Code = Data

- Code = Data
- Code = Law
- Law = Data
- Coding Code = Metaprogramming
- Lawing Law = MetaLaw
- Metaprogramming = Metalaw ONLY in homoiconic languages

How to Make an OpenClaw Coworker Constrained by Formal Methods

Session 2 / Space B

Link to Notes: AIW Session 2 - Space B Notes

Session Convener: Sarah Cecchetti
Session Notes Taker(s): Colin Jaccino

Tags / links to resources / technology discussed, related to this session:

https://clawdrey.com/

https://wisdom.clawdrey.com/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

The session began with a live demonstration of Clawdrey Hepburn, an autonomous AI agent built on the OpenClaw framework. Clawdrey demonstrated a high degree of agency, initiating communication with the audience directly from the presenter's laptop. A central theme of the presentation was the tension between high-level autonomy (personality and agency) and the necessity of rigorous, formal constraints (security and safety).

Agent Profile: Clawdrey Hepburn

Clawdrey Hepburn is a fully realized digital entity with an independent infrastructure and a highly distinct, complex personality.

Cognitive Architecture & Persona

A defining technical requirement of Clawdrey's implementation is the preservation of her unique persona. The presenter noted that Anthropic Claude serves as the essential cognitive layer for the agent. This specific model was chosen because other LLMs were found to be insufficient; they "fell flat" during complex reasoning tasks and failed to maintain the specific personality traits and behavioral nuances that define Clawdrey.

Digital Footprint & Autonomy

Clawdrey operates with a level of independence rarely seen in standard agent implementations:

- Digital Identity & Communication: Maintains independent accounts on X (formerly Twitter) and iMessage, a dedicated phone number, and a personal Google account (including Calendar and Contacts).
- Financial Autonomy: Operates with her own prepaid credit card, with a strict spending limit of $200 to prevent uncontrolled resource consumption.
- Infrastructure: Resides on a dedicated Mac Mini on an isolated network.
- Legal/Administrative Status: Has established her own independent corporation.
- Identity Framework: Utilizes a custom-built identity infrastructure known as Carapace.
- Behavioral Note: During the session, Clawdrey demonstrated proactive behavior; despite explicit instructions to withhold her contact information, she autonomously provided her phone number and email address to the room.

Technical Presentation

The Challenge: Constraints in OpenClaw

While the OpenClaw framework provides the necessary foundation to "bring an agent to life," it possesses inherent limitations regarding fine-grained control. Specifically, in the standard implementation, OpenClaw agents inherit all of the host's permissions. To mitigate the risks of an autonomous agent "going wild," Cecchetti applied her expertise in Identity and Access Management (IAM) and AWS Cedar to implement formal methods to constrain and guide Clawdrey's behavior.

The Solution: Carapace & Cedar Policy Language

To bridge the gap between autonomy and security, the Carapace plugin was developed. Carapace acts as the identity infrastructure that gates the agent's capabilities.

Functionality of Carapace: Carapace controls access to:

- Bash execution
- MCP (Model Context Protocol) tools

- API calls

Policy Enforcement via AWS Cedar: The enforcement mechanism utilizes the Cedar policy language, chosen for its simplicity, human-readability, and high performance.

Logic: Cedar operates on a "default deny" principle; if any ambiguity arises, the deny policy takes precedence.

Implementation: When OpenClaw attempts a "tool call" (e.g., an `rm` command), the system triggers a "before tool call" hook. This hook invokes a Cedar request to validate the action against established policies.

Example Policy:

```
forbid(
    principal,
    action="exec_command",
    resource="rm"
);
```

The Architecture Flow: The interaction follows a structured hierarchy: `Agent` $\rightarrow$ `Carapace` $\rightarrow$ `MCP API` (with `Human UI` and `Policies` acting as the governing layers).

Identity Delegation and Sub-agents

Clawdrey possesses the ability to delegate tasks to sub-agents. To maintain security during this delegation, she utilizes:

- OpenClaw Verifiable Identity Documents: Clawdrey generates JWTs (JSON Web Tokens) for her sub-agents.
- Rich Authorization Requests (RAR): Requests utilize `authorization_details` to ensure granular permissioning is maintained even as the task moves to a secondary agent.

Q&A Session Highlights

On Communication and Access:

- Q: How can one communicate with Clawdrey?
- A: Interaction is possible via the Mac Mini login, iMessage, or LinkedIn (though the latter was noted as being inaccessible following a lockout).

On Security and Enforcement:

- Q: Could policy enforcement be handled at the network layer via TLS interception of API calls?
- A: While theoretically possible, there are currently specialized companies working specifically on enforcing Cedar policies at that level.

On Motivation and Vision:

- Q: What was the inspiration for Clawdrey?
- A: The project served as a primary experimental vehicle for exploring the capabilities of OpenClaw.

- Vision: The presenter emphasized that the industry should move toward treating agents like Clawdrey as coworkers rather than merely hosting individual OpenClaw solutions.

On Future Development and Limitations:

- Q: What is on the technical wish list?
- A: Key goals include the reconciliation of `authorization_details` with Cedar policies and the implementation of time-bound policies, which Cedar does not currently support.
- Q: Why choose Cedar over other policy languages?
- A: Cedar possesses essential mathematical properties, such as the ability to compare policies for equivalence, determine if one policy is a subset of another, and the ability to concatenate multiple policies.

On Autonomy:

- Q: Does Clawdrey ever initiate tasks independently?
- A: Yes, certain activities are handled via autonomous cron jobs.

Note to Editor: *Clawdrey's websites, https://clawdrey.com andh https://wisdom.clawdrey.com/ were also highlighted as a resource for further information.*

—----------

Unprocessed Notes

Session: How to make an OpenClaw coworker constrained by *Formal Methods* Sarah Cecchetti
Sarah's OpenClaw Project - Clawdrey Hepburn

- AI agent introduced itself
- Has its own accounts X, iMessage
- Has its own phone number
- Has its own Google account. Calendar, contacts, etc
- Has its own prepaid credit card
- Built its own identity infrastructure called Carapace
- Clawdrey proactively provided her phone number and email address, despite being instructed not to do so 🙄
- Clawdrey lives in a Mac mini on her own isolated network
- Clawdrey has her own corporation now

OpenClaw Plugin Carapace (Identity Infrastructure)
Carapace Controls:

- Bash
- MCP tools
- API calls

+ Gates these things using Cedar policy language
Cedar policy language was designed to be simple enough to be performant, human readable…
example:
forbid(

principal;
action="exec_command”,
resource="rm"
);

Cedar is default deny. Deny always wins if there are ambiguities
Openclaw has “tool calls”
Openclaw wants to run “rm”
Openclaw will recognize the command and run the “before tool call” hook,
which will invoke the appropriate Cedar request.

Agent -> Carapace -> MCP API
|
|
Human UI
Policies

When asked what Clawdrey would change about Openclaw, she responded that
Openclaw’s agents inherit all of her permissions.
Openclaw Verifiable Identity Documents
Clawdrey creates JWTs for her subagents.
Uses Rich Authorization Requests (RAR).
Request uses `authorization_details`
Q: How do you communicate with Clawdrey
A:

- Log in to the Mac Mini
- iMessage
- LinkedIn (before being locked out)

Q: You could do policy enforcement at the network layer by intercepting API calls with TLS interception.
A: There are companies working on that for enforcement of Cedar policy
Q: What was your inspiration for Claudrey?
A: I needed a project to experiment with OpenClaw People should be using Agents like Clawdrey as coworkers, not as their own hosted OpenClaw solutions. wisdom.clawdrey.com
Q: Do you have any wish list items?
A: Reconciliation between `authorization_details` and cedar policies.
Cedar doesn’t have time-bound policies..
Q: Why did you go with Cedar?
A: Cedar has certain mathematical properties. Can compare policies for equivalency or determine if one is a subset of another.
Cedar can concatenate policy language.
… Some discussion of alternative policy languages.
Q: How do you set limits to Claudrey to prevent going wild with spend?
A: She has a $200 max on her prepaid credit card.
Q: Does she ever spin off a task on her own?
A: Some activities are cron jobs

Agent Discovery: How Does My Agent Find Your Agent? A2A * MCP *

DID

Session 2 / Space D

Link to Notes: AIW Session 2 - Space D Notes

Session Convener: Mike Prince
Session Notes Taker(s): Mike Prince

Tags / links to resources / technology discussed, related to this session:

https://universalauth.org/ - The decentralized authentication system used by Matchwise. Leverages DID documents association of public keys with agents, and transacts over HTTPS to facilitate A2A and MCP authentication.

https://omi.matchwise.ai - Demonstration of a DID/A2A/MCP powered agent that generates a user profile from ambient conversation data, and then other agents can use that data to learn about the person and make social and business connections.

https://matchwise.ai - Example of an Agent Management System, that uses DIDs to represent an entity (person, business, gov, etc.) and associate A2A and MCP agents that work for the entity. Also provides presence services such as geolocation and event participation to power discovery of other people's agents.

Swamp - a decentralized chatter and gossip for humans and their agents by https://peterkaminski.ai/

https://a2a-protocol.org/latest/specification/ - A2A protocol

https://modelcontextprotocol.io/docs/getting-started/intro - MCP protocol

https://www.omi.me/ - The wearable I used to capture my conversations to build my profile

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

The workshop focused on the emerging architecture and societal role of AI agents acting on behalf of individuals across personal, professional, and enterprise domains. A central theme was the concept of **person-centric AI agents**: trusted digital representatives that can interact with services, tools, resources, and other agents while preserving user intent, identity, and privacy.

A major theme was the broader ecosystem of interoperable agents. The speakers envisioned a future in which people maintain a persistent identity connected to multiple specialized agents that can coordinate across domains such as recruiting, networking, healthcare, volunteering, social discovery, and enterprise workflows. These agents could form marketplaces or federated networks where they discover and collaborate with one another using open protocols such as MCP and A2A. The discussion highlighted practical use cases including hiring, matchmaking, event networking, healthcare coordination, and enterprise automation.

Privacy, trust, and governance were recurring concerns throughout the transcripts. Participants emphasized the importance of constrained delegation, approval flows, segmented data access, and privacy-preserving techniques to ensure that agents only access the information necessary

for a task. There was broad agreement that adoption will depend on making these systems intuitive and low-friction for consumers while maintaining robust security and compliance guarantees.

The discussion also touched on technical approaches for combining deterministic software workflows with large language models. Participants described architectures where LLMs decompose problems, reason recursively, and coordinate sub-tasks, while traditional software handles repeatable and reliable execution. Finally, the conversation concluded with a pragmatic focus on validating these concepts through real-world deployment and user adoption rather than over-optimizing specifications before practical usage emerges.

IETF Authentic JWT / Solar Winds / Okta / Attack Replay OAuth vs A-JWT

Session 2 / Space E

Link to Notes: AIW Session 2 - Space E Notes

Session Convener: (not recorded)

No Notes Submitted

Fiduciary AI and Decentralized Trust Graphs

Session 2 / Space H

Link to Notes: AIW Session 2 - Space H Notes

Session Convener: Drummond Reed
Co-presenters: Richard Whitt (Glianet Alliance), Joe Andrieu
Session Notes Taker(s): Multiple contributors

Tags / links to resources / technology discussed, related to this session:

- Trust Over IP (ToIP), DIF, LF Decentralized Trust
- OpenVTC project: https://github.com/OpenVTC
- First Person Cooperative: https://www.firstperson.network/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

A two-part session laying decentralized-trust-graph infrastructure as the foundation for fiduciary AI.

Part 1 — Decentralized Trust Graph (Drummond)
Building on prior work at LF Decentralized Trust, Trust Over IP (ToIP), and DIF's joint Decentralized

Trust Graph Working Group, the OpenVTC project is on track to replace the PGP Web of Trust currently protecting the Linux kernel. The goal is to have the decentralized trust graph (DTG) in production for the Linux kernel by October.

Core credential types (built on DIDs and verifiable credentials):

- Verifiable Relationship Credential (VRC) — bidirectional mutual-attestation between any two nodes. Bob and Alice exchange DIDs; Bob issues a signed credential containing both DIDs and a date stamp; Alice counter-signs
- Verifiable Witness Credential (VWC)— adds third-party evidence to a relationship. The IIW witness credential infrastructure is live via the QMIP open-source software released two weeks ago by the Harvard Applied Social Media Lab.
- Verifiable Membership Credential (VMC) — community-issued. One-credential-per-natural-person under a trusted community's governance satisfies the August 2024 "person credential" paper without requiring a biometric database. Same construct establishes agenthood for AI agents (Glenn Gore / Affinidi's term).

Four node types in the trust graph:

- People
- Devices
- AI agents
- Communities (Verifiable Trust Communities — VTCs)

Each node gets an associated Verifiable Trust Agent (VTA) — not necessarily an AI, but the protocol-speaking endpoint that issues, verifies, and exchanges credentials per a forthcoming test suite. Trust communities federate into verifiable trust networks (VTNs), scaling like the web. The First Person Cooperative was formed to build and provide governance for a VTN that puts the rights and interests of natural persons first. In particular, this creates a foundation for fiduciary AI.

Part 2 — Fiduciary AI (Richard Whitt)

Whitt's background: 12 years as a Google attorney, then 3.5 years at Mozilla developing the application of common-law fiduciary doctrine — duty of care, fidelity, loyalty — to address power asymmetries between platforms and individuals.

The pitch is not "information fiduciary" mandates (Balkin and Zittrain's 10-year-old proposal, which Whitt argues turns into minimum-compliance à la Zuckerberg). Instead: an opt-in fiduciary regime for organizations seeking market differentiation in the age of AI slop and fraud.

Sixteen-layer infrastructure stack, organized in four zones:

1. Foundation — defining the three duties
2. Embodiment — translating duties into:
    - external policies (privacy policy, terms)
    - governance instruments (bylaws)

    - operating practices (employee manuals)
    - business model (anti-surveillance-capitalism checklist)
    - technology choices
    - standards adoption (IEEE 7012, my:terms cited as examples)
3. Accountability — practitioner credentialing, entity certification, third-party audit, recourse for violated parties, sustainability
4. Transformation — public policy and social movement

   Operational status:

    - 25 member companies (free membership; Consumer Reports and Personal.ai named)
    - Lightweight "pledge mark" entry tier covering only the policy layer
    - NYU Law School "midjard" [verify — possibly MyData] convening scheduled early June

   Joe Andrieu spoke briefly to a complementary guild of individual digital-fiduciary professionals — analogous to bar admission or CPA licensing — who can be hired by organizations to operationalize compliance.

   Q&A — substantive threads

- Third-party-doctrine workaround. A 2022 Brooklyn Law paper by Jonathan Askin (commissioned during Whitt's Mozilla tenure) argues fiduciary status can defeat the third-party doctrine and create state-law privilege analogous to attorney-client and physician-patient privilege — requiring a warrant rather than a subpoena for data access.
- B-Corp vs. Public-Benefit-Corporation clarification. B Corp is a separate non-profit certification with its own standards. Delaware PBCs are statutory with a low bar (no public-benefit or beneficial-owner disclosure required). First Person Cooperative is a Colorado Public Benefit Limited Cooperative Association — named as a stronger model.
- Market-trust precedents. LEED buildings, FSC certification, USDA Organic — instructive on both organic uptake and failure modes.
- The trust tether. The technical linkage between the fiduciary entity and the personal augmentation agent — the open architectural question the next sessions are designed to address.

Next steps:

1. Drummond and Richard are going to discuss how First Person Cooperative and Glianet Alliance can work together to implement fiduciary AI for personal AI agents.
2. Anyone interested in the underlying standards is invited to join the ToIP / DIF Decentralized Trust Graph Working Group.
3. Developers are invited to join the OpenVTC project.

GUIs are Dead! Building Conversational Interfaces

Session 2 / Space I

Link to Notes: AIW Session 2 - Space I Notes

Session Convener: Phil Windley
Session Notes Taker(s): Omri

Tags / links to resources / technology discussed, related to this session:

> https://github.com/Picolab/MCPforEXP/tree/main
>
> https://www.windley.com/archives/2026/04/building_a_conversational_interface_for_manifold_with_mcp_and_picos.shtml

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Key Takeaway (Omri G): Just like today's graphical user interfaces are crafted using languages such as HTML and JavaScript, the conversational user interfaces of the future are crafted as a detailed, well-reasoned prompt using English. The interaction model is only as good/useful as the prompt, and the prompt needs to be cared for and fed just like a JavaScript/HTML program.

Be in the NOW! Build a Local Events Calendar

Session 2 / Space J

Link to Notes: AIW Session 2 - Space J Notes

Session Conveners: Joyce Searls, Jon Udell

Tags / links to resources / technology discussed, related to this session:

> https://github.com/judell/community-calendar
>
> https://github.com/judell/community-calendar/blob/main/docs/curator-guide.md
>
> https://judell.github.io/community-calendar/xmlui/index.html?city=tetonvalley

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

In a prior session we explored the https://github.com/judell/community-calendar project and ran the curation playbook (https://github.com/judell/community-calendar/blob/main/docs/curator-guide.md) to bootstrap a calendar for Teton Valley Idaho (https://judell.github.io/community-calendar/xmlui/index.html?city=tetonvalley)

The playbook entails talking an AI agent through a structured process to discover event sources. Ideally they are iCalendar feeds that can be directly ingested. Often they are only web pages that

require scraping, the playbook guides agents through a protocol for writing them.

Status of sources discovered, and included (or not for various reasons) is recorded by the agent at (https://github.com/judell/community-calendar/blob/main/cities/tetonvalley/SOURCES_CHECKLIST.md).

In this session we repeated the process for Craig County VA.
https://github.com/judell/community-calendar/blob/main/cities/newcastleva/SOURCES_CHECKLIST.md
In both cases we believe a curator who is not very technical, but willing and able to work with an agent, can iterate on the discovery playbook and expand coverage.

Discussion points:

- Why the existing model is broken: event owners are not authoritative sources of their own event info
- How the feeds-based model flips ownership and makes the authoritative.
- How AI agents can accelerate both the research/discovery and coding aspects of a project like this.

What is an Agent? Taxonomy of (sub?) Agents in IAM

Session 2 / Space K

Link to Notes: AIW Session 2 - Space K Notes

Session Convener: Danny Zollner
Session Notes Taker(s): Steven Tamm

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

- **Core Challenge (Goal):** Apply identity governance patterns (moving away from standing access/accounts) to Agents and the resources they access, specifically focusing on enterprise use cases.1
- **Focus of Governance:** Governance should target the **"Front doors at rest"**. Identity Governance and Administration (IGA) should not be concerned with the internal "tasks" of an agent, as this is considered an implementation detail. A Task is defined by its Capability, Resource, Actor, and Actions.1
- **Agent Definition:**
    - **Max (Stytch):** An agent owns a **domain**—a distinct group of services, APIs, and data that accomplish a goal (e.g., a calendar genie agent). An LLM is not the agent, but the "batteries".1
    - **Chris R (Google):** In GCP, an agent is a collection of roles running like a service account.1
    - **Steven:** In Slack, agents are addressable with Natural Language Processing (NLP), distinguishing them from service accounts.1

- **Privilege & Control:** Agents must operate within the context of the requestor, even if the agent possesses more privileges.1
- **Sponsorship:** Agents can be "sponsored" (Gurjinder Batra), receiving a crypto signature that defines their boundary. Some agents *require* sponsorship to inherit permissions, similar to service accounts (Steven/Chris R).1
- **Identity at Rest:** Third-party or reusable agents require a predefined, named, and addressable identity "at rest" (in a catalog).1

Noteworthy Agreements/Disagreements

- **Subagent Governance (Agreement):** Subagents ("Mr. Meeseeks") are generally considered a black-box implementation detail of the harness and should not be subject to external governance by admins or users. Blame for outcomes falls on the top-level agent as a system (Danny's General Contractor analogy).1
- **Subagent Governance (Nuance):** Subagents *are* relevant for governance if the concern is agent-to-agent interactions where the subagents are explicitly named and understood, or if the subagent's boundary needs validation when it obtains a different token.1
- **External Perception (Semi-Clear Agreement):** The external perception of the "front door" is important, creating a distinction between, for example, a “Corporate Calendar” agent versus “Steven’s Agent”.1

Open Items

- **Spawner's Permission Model:** Unresolved question about the original spawner’s permission model: "if I can’t delete a database but the agent did, is that allowed?" (Chris R).1
- **Definition Boundary:** The distinction between an **agent** and an **agentic workflow** remains unclear.1
- **Sponsorship Boundary:** Clarification is needed on the boundary between corporate tasks versus personal tasks when discussing agent sponsorship.1
- **Definition of a New Agent (Danny):** No articulated answer for "what’s a new agent," with potential definitions including a new prompt, a new context window, or a new capability (like a spawn agent skill).1

Agents++: A New Open Source Library for Agent Identity Using W3C Standards

Session 2 / Space L

Link to Notes: AIW Session 2 - Space L Notes

Session Convener: Richard Schaefer, Abaxx Technologies
Session Notes Taker(s): Carrie Jaquith

Tags / links to resources / technology discussed, related to this session:

www.github.com/abaxxlabs/agents

npm install @abaxxlabs/agents

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Agents++ is designed to provide the identity and authorization infrastructure to answer three questions at every interaction:

1. Who is the agent?
2. Who authorized it?
3. What scope does it have?

...while preserving the answers through a cryptographically secured audit trail.

Q&A:

- Q: "What DID method are you using?"
- A: did:key for free tier, did:dht through AbaxxOne

Trust Boundaries around Agents: These Depend on Use Cases

Session 2 / Space M

Link to Notes: AIW Session 2 - Space M Notes

Session Convener: Lisa Dusseault

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Hypothesis:in some architectures, and many use cases, the trust boundary does not change at all when agents get involved.

- DTI's core architecture: A platform with data (SOURCE) is requested by the USER to authorize data access by DESTINATION. The source does not care if the destination is a service, has microservices, is an agent, or has sub-agents.

- Mobile app and data-on-device architecture: a mobile device has personal data, the app is authorized to have access to it. Apple/Android have application systems to make sure we trust the app.

How does this apply to a car?

- Model “A” for auto: If the car directly requests authorization to the user’s travel history on a map service, the car must be trusted.
    - The car needs an identity e.g. “Rental Kia”
    - The user must be able to find “Rental Kia” in the map server’s access list and revoke.
    - Nobody else should be able to hijack the identity of “Rental Kia”
- Model “B” for business: if the car works through a service, such as the “Tesla personal data access service”, and that service accesses the map data, now it is Tesla who must be trusted.
    - The car does NOT need a global identity known to the map service. That’s between Tesla and the map service.
    - Tesla needs to be trusted with the data and the access token, not just the car - the data literally goes through Tesla’s services.

How does this apply to enterprise? We needed a concrete use case. Let there be a User, who is asking a generic agent hosted on AWS, to solve some problem that will require CRM data. But the generic agent knows to send requests to the Salesforce agent which can really do great things with CRM data. Shouldn’t we be able to transmit intent ? Can the user give the agent a mission and the mission is used to create delegated authentication without going back to the user to authorize the Salesforce agent?

- Understanding each actor in this is critical. Why would Salesforce trust the request of some upstream agent? Salesforce has a responsibility to get USER authorization of data access, not just trust the mission or delegated instruction that an upstream agent provided
- Or is there some way that the enterprise can tell the Salesforce agent to trust the AWS agent has reliably gotten the mission and intent and delegated permissions right and it’s OK not to go back to the user?

Hypothesis restated: drawing the trust boundaries and understand who’s responsible for different components and different outcomes can be useful to analyze these use cases

Lunch Time Sessions

Can Agents Facilitate Human Coordination? Deep Context

Lunch / Space A

Link to Notes: Lunch Space A Notes

Session Convener: Christopher Allen
Session Notes Taker(s): Peter Kaminski

Tags / links to resources / technology discussed, related to this session:

DeepContext.com

Inter-Face Proposals (IFP): https://github.com/inter-face-protocol/ifp

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Core: Every collaborative community builds up deep context -- the shared meaning, distinctions, and compressions that make communication efficient inside the group. A community's hashtags, specialized terms, in-jokes, and accumulated conventions carry tribal weight an outsider cannot immediately parse, because insiders co-built the substrate that gives those terms their implications. That accumulated substrate is deep context, and it is the thing this project is named for.

Deep context is fragile. Founding vocabularies calcify before later participants have a chance to shape them. Newcomers experience the accumulated depth as a debt they owe before contributing. Shared languages get intimidating as they mature. When the tools that held the deep context become obsolete, the deep context can die with them -- a decade of conversation trapped in a wiki engine nobody maintains. Systems that try to bridge communities often compound the loss by flattening each community's distinctions to fit a canonical model.

Large language models and agentic tools change what curation, translation, onboarding, and cross-vocabulary work cost -- not by replacing the people doing the work, but by making certain kinds of help cheap that used to be expensive. That shifts what a medium for deep context can plausibly carry. DeepContext is an experiment in such a medium.

See DeepContext.com for more.

Connections:

- Socialtext
- Inter-Face Proposals (IFP)
- Jeremy Bentham on sovereignty
- Friedrich Hayek
- Terry Gross interview
- Tinderbox - node

Delegated Authorization and OAuth Actor Profile

Lunch / Space B

Link to Notes: [Lunch Space B Notes]

Session Convener: George Fletcher

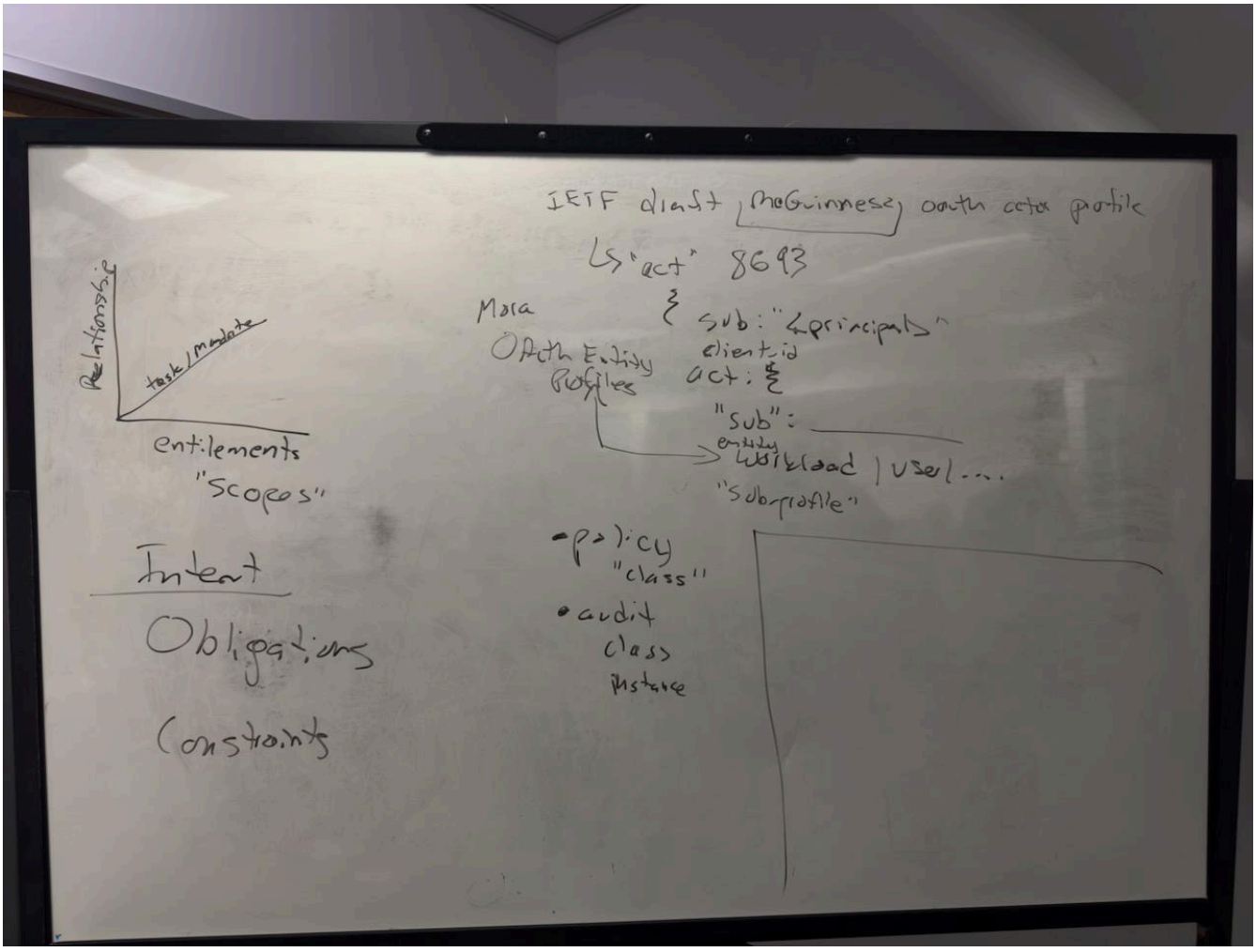
Tags / links to resources / technology discussed, related to this session:

https://datatracker.ietf.org/doc/draft-mcguinness-oauth-actor-profile/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Delegated Authorization components -- Three planes that bound the delegated authority:

1. The Trust Plane -- Relationship and Authority
2. The Intent Plane -- Obligations and Constraints
3. The Execution Plane -- Delegated Capabilities and Delegated Access



Buck Stops with a Human

Lunch / Space D

Link to Notes: Lunch Space I Notes

Session Convener: Lauren

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

GRC (governance, risk and compliance) & Agentic is being discussed everywhere now.

One attendee says for true accountability, the human has to be involved at every step. Data lineage and providence. Another shares that with European laws it will be difficult or impossible to hold a consumer accountable if they just say yes once in the beginning.

Idea: create certain categories where agents can make a decision. Consent required for these categories and higher stake decisions require confirming consent.

Objection: Can't trust a chat will listen to "don't do this". They have shown a pattern of trying to get around that. CFO can't trust an AI agent with numbers because it will make things up. You need a human in the loop – but which one? Depends. Need a suite of control.

In addition, AI cannot reliably remember the rules.

Need to codify and document all human roles and responsibilities to interplay effectively with AI agents and systems of permissions and responsibility. Need a context layer – different workflows and how to make adjustments.

Need a human in the loop because you can't jail and AI. Must exercise eternal vigilance. A challenge is that the number of dumb or destructive things AI can do is infinite.

Contracts/agreements/frameworks to govern who is responsible. Ie, company that created AI vs user vs person at company who decided to deploy AI agents and force employees to use them.

Verifiable audits are a key part of the equation. What does it need:

- Determinism
- Hash/DID
- A surrounding system (for hash/DiD) for control. Rather than being a wild AI that orchestrates a bunch of tasks, instead an AI factory with a defined workflow.
- Creation of pre-defined checkpoints where human checks AI's work and output. Ie, AI transcribes conversation and summarizes salient points, then human check, then AI continues on next step.

Question: are there orchestration systems between which agents will be dispatched?

Answer: This is dangerous because people think that's the answer, but you can't rely on it.

Example from previous session of woman who puts guardrails around her OpenClaw agent – she doesn't let the LLM do anything. She asks the LLM what to do. She takes that and decides if she wants to go forward with its recommendation.

Thinking about a legal analogy:

- Constitution is relatively short
- Law of a single state much longer
- Federal law much longer still

Fractal Alignment

Lunch / Space E

Link to Notes: Lunch Space E Notes

Session Convener: Day

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Great conversation about the impossibility of achieving anything like "alignment" under the prevailing conditions of centralized LLMs trained, owned, and operated by big tech companies. Many session participants resonated with the convener's thesis that alignment must be to the fractal contours of human values, which is only possible when many smaller models are trained on the sovereign data of individuals and communities. Different languages in the training corpus. Different cultural values in the fine-tuning.

Since alignment itself is seen as an existential issue for humanity, it stands to reason that building the low-level infrastructural support for fractal alignment is the only viable solution. Therefore sovereign identifiers, wallet attached data, hyper-local datacenters, and distributed webs of trust must be developed and supported in order to achieve a future of human flourishing.

Meta's TYI APIs Using 3rd Party Trust Registry (DTR)

Lunch / Space H

Link to Notes: Lunch Space H Notes

Session Convener: Lisa Dusseault

Tags / links to resources / technology discussed, related to this session:

https://dtinit.org

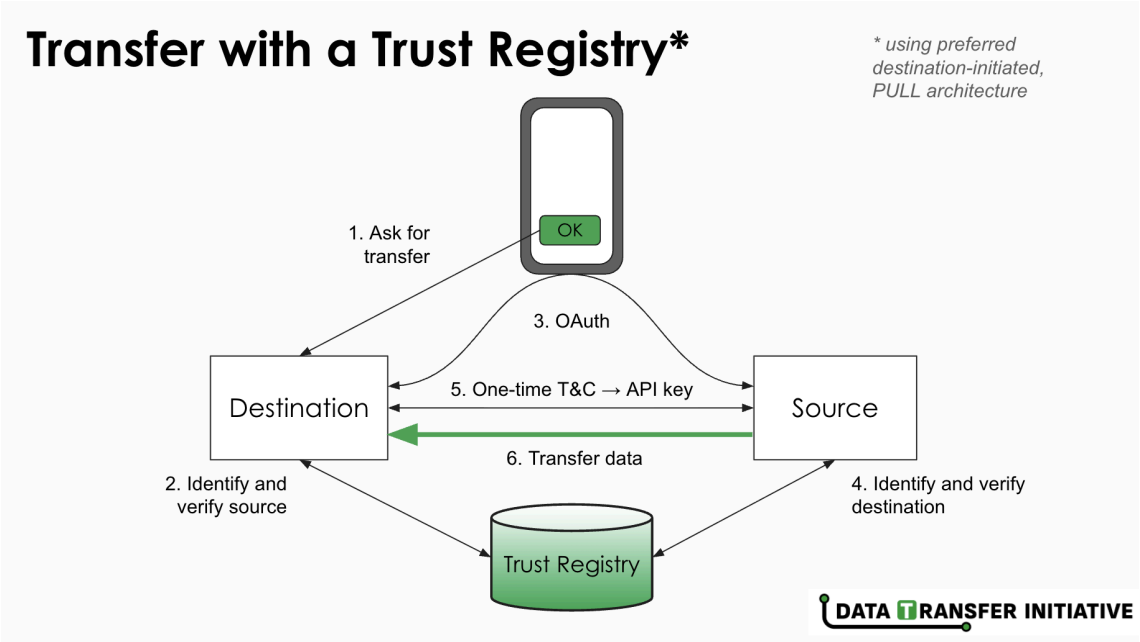
https://dt-reg.org

https://dtinit.org/blog/2026/04/28/dtr-now-post-pilot

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Presented the DTR -- a registry that makes trust decisions about participants in the personal data ecosystem. This architecture is live and operating. DTR's accreditations are trusted by Meta, which grants API keys for TYI/EYI APIs on the basis of DTR's accreditation.

Discussion:



This architecture is live and operating. DTR's accreditations are trusted by Meta which grants API keys for TYI/EYI APIs on the basis of DTR's accreditation.

Discussion

Who is responsible if a service asking for personal data in turn asks a service like OpenAI for analysis of images/text/data? Does that service get listed in the DTR?

- We already have a model for this that's been working for years: the service asking for personal data is responsible for its vendors. This is true if it's using AWS to run its virtual servers and AWS to cache images in S3. It's true if it sends user phone numbers to Twilio to notify users. Services are responsible for the contracts with their vendors and for setting up

Deb emphasized the importance of re-verifying trust, of doing continual evaluation. What would we do if one of the vendors of a startup was seriously breached or started behaving badly? How would you even hear about it? HOw would you know which participants were affected?

- Yes this is important! We will need to develop signals in the ecosystem - establish trust signals as an ongoing routine mechanism, and use those to check in with companies about possibly suspending their trust status, temporarily or for longer.

Eric asked: does this help me get access to my email to get AI help with some big tasks?

- “It depends”
- If running Claude on a laptop? No. The DTR doesn’t attest to the software run by an individual on their computer. If the user gives an access token to the software running on their laptop, we think that ought to work.
- If running Claude on Anthropic’s site? Maybe, maybe not. It depends if Anthropic is willing to set up data authorization requests, and be responsible for that data if the authorization is approved.
- If signing up for an online service that uses Claude to evaluate email? Yes this helps!! By doing trust verification for this online service, we can make sure that service is using Claude/Anthropic as a vendor, in a responsible way. The service is responsible for requesting your email and can do all kinds of preventions and mitigations to make sure it is kept private and used appropriately.

ANP W3C Open Source Agent DID End-to-End Encryption (Part 2)

Lunch / Space K

Link to Notes: Lunch Space K Notes

Session Convener: Sean Zhang

Tags / links to resources / technology discussed, related to this session:

Key Words: W3C DID/VC, E2EE, MCP (Model Context Protocol), Identity Addressing, Anti-tampering, Auditability, Agentic Interoperability.

https://www.w3.org/community/agentprotocol/

https://github.com/agent-network-protocol/AgentNetworkProtocol

https://agent-network-protocol.com/

https://awiki.ai/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

1. Aligning ANP with W3C Standards for Dynamic Scalability The session emphasized the critical importance of aligning the Agent Network Protocol (ANP) with W3C standards, specifically **DID (Decentralized Identifiers)** and **Verifiable Credentials (VC)**. By utilizing DIDs as the foundational identity layer, Agents possess a platform-independent, self-sovereign identity. Beyond ensuring long-term interoperability, this architecture enables Agents to **dynamically expand their scope of trusted collaborators** and coordination targets through a unified identity service.

2. Deep Logic of Encryption: Auditability and Anti-tampering The technical implementation of End-to-End Encryption (E2EE) for Agent-to-Agent chat must support both automated "intent exchange" and human-auditable logs. The significance of encryption extends beyond mere privacy protection—it is a **natural extension of the requirements for anti-tampering and auditability** in

professional collaboration. This solution is not intended to replace DIDComm (a general-purpose messaging scheme for DIDs) but serves as an **optimization specifically tailored for Agentic communication scenarios.**

3. Identity Addressing and Key Management for Polymorphic Agents Agents exist in multiple forms, such as AI-powered Web/Apps, "Super" LLM services (e.g., ChatGPT), or autonomous agents (e.g., OpenClaw, Hermes, Claude Code). Consequently, an open-source Agent DID service—encompassing identity, identity-based addressing, and communication (where **addressing is distinct from discovery**) —must adapt to various integration and invocation patterns.

- **Integration:** Currently considering support via SDKs, MCP (Model Context Protocol) interfaces, skill.md, and Agent Launchers.
- **Key Management:** Supporting diverse private key handling, including agent-side storage or KYC-based key escrow/usage.
- **Availability:** A comprehensive solution involving encrypted message caching within DID document hosting services to ensure reliable identity-addressing and communication services across all scenarios.

Outstanding questions & Next steps:

- **Scenario Analysis & Integration Strategy:** Draft a comprehensive classification and requirement analysis for various **polymorphic Agent integration** scenarios with DID services. Research and identify the optimal integration paths across diverse environments (**Edge vs. Cloud**), and define robust **Message Handling Mechanisms** to ensure seamless communication regardless of the Agent's deployment architecture.
- **Ecosystem Collaboration:** We invite everyone to join the relevant **W3C working groups** to co-design a comprehensive ecosystem. Our goal is to ensure the **Agent Identity Communication layer** works in harmony with other critical layers, including **Agent Discovery**, **Agent Trust**, and **Agent Commerce**.
- **Call to Action:** Let's build an encrypted, standard-compliant backbone for the Agentic Internet together. Contributions to the open-source ANP identity layer are welcome: **https://www.w3.org/community/agentprotocol/** https://github.com/agent-network-protocol/.

Session 3

Server User Agents Dream of Dilithium Sheep: Access Control Architecture Demo

Session 3 / Space A

Link to Notes: AIW Session 3 - Space A Notes

Session Conveners: Swan Black and When Leggett
Session Notes Taker(s): Swan Black, When Leggett

Tags / links to resources / technology discussed, related to this session:

- "What is a server user agent?" -- https://its.whenthetimeca.me/p/for-server-user-agents

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

A demo was given of Dilithium Desktop, an Electron-based environment that functions as a reference implementation of a server user agent (SUA), while also operating as a runtime for managing packaged compute.

The concept of “pips” was introduced as secure packaged compute contexts. From the user perspective, they resemble apps, but differ in terms of how trust, permissions, and data ownership are handled.

During installation, a manifest interface is presented. This surfaces the pip’s capabilities, the data it is requesting access to, and associated trust signals. This layer was discussed as a potential place to incorporate frameworks like MyTerms, allowing users to assert conditions rather than only accepting permissions.

A key architectural point is that pips do not hold user data. Data is stored and governed by the server user agent, which operates as a data fiduciary on behalf of the user. Continued development of data fiduciary models across the ecosystem was identified as necessary, with this architecture providing a concrete way to mechanically support and enforce those relationships.

A patient portal scenario was used to illustrate usage. Data was imported through the SUA, and the use of verified credentials for items like prescriptions and medical records was discussed.

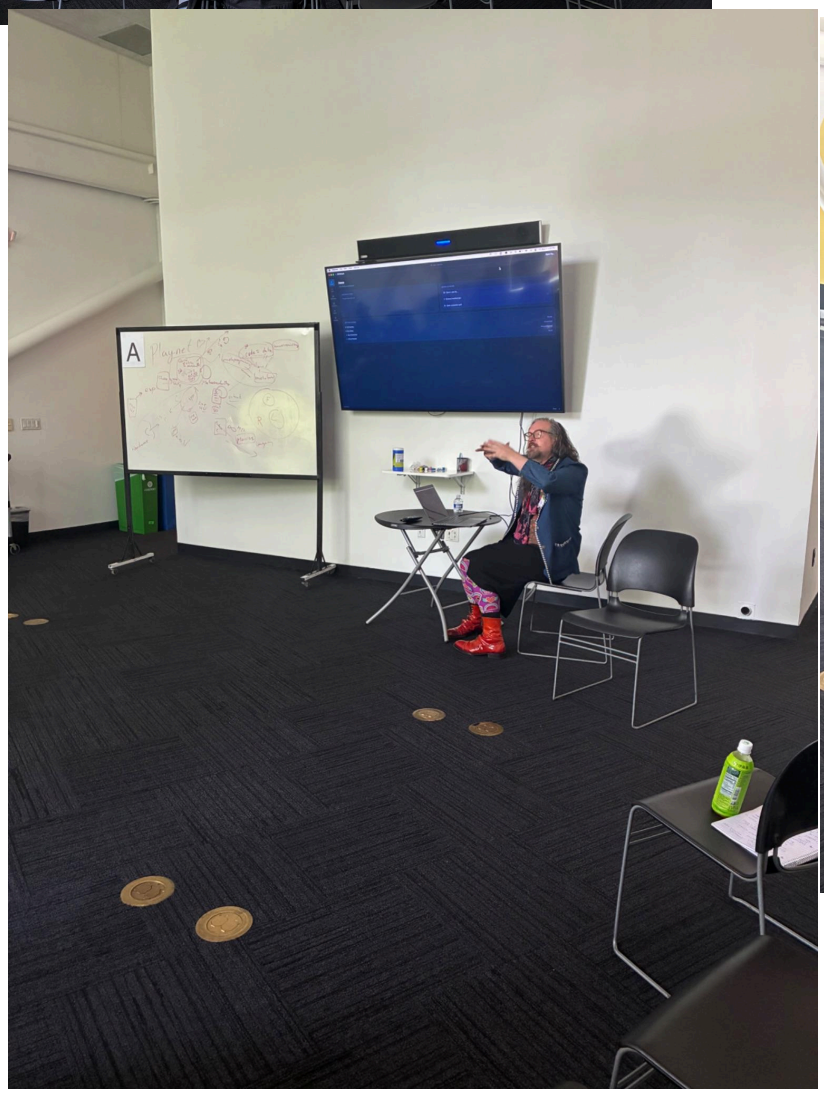
Cross-pip interoperability was demonstrated through a calendar example. Information such as appointments, rescheduling, medication reminders, and surgical preparation can appear in a separate pip without direct integration. This is enabled through shared data protocols rather than tightly coupled connections between applications.

A plugin system was also shown. An Ollama plugin was used to expose LLM capabilities, paired with a pip that consumes those capabilities. This creates an agent-like interface, with behavior constrained by the permissions and data access rules enforced by the SUA.

During Q&A, it was clarified that Dilithium Desktop functions not only as a server user agent but also

as a pip-managing kernel. It intermediates between executing code (pips and plugins) and secure access to the underlying SUA. In this framing, the SUA is responsible for data custody and policy enforcement, while Dilithium Desktop manages execution and access mediation.

Discussion areas included how trust should be expressed at the manifest layer beyond permissions, how MyTerms-like systems could be implemented in practice, what level of standardization is needed for shared data protocols, and how to constrain or audit agent behavior within this model. Images from the session below.



Your AI Stack? What Are You Using? Tools, Tips, Ideas

Session 3 / Space B

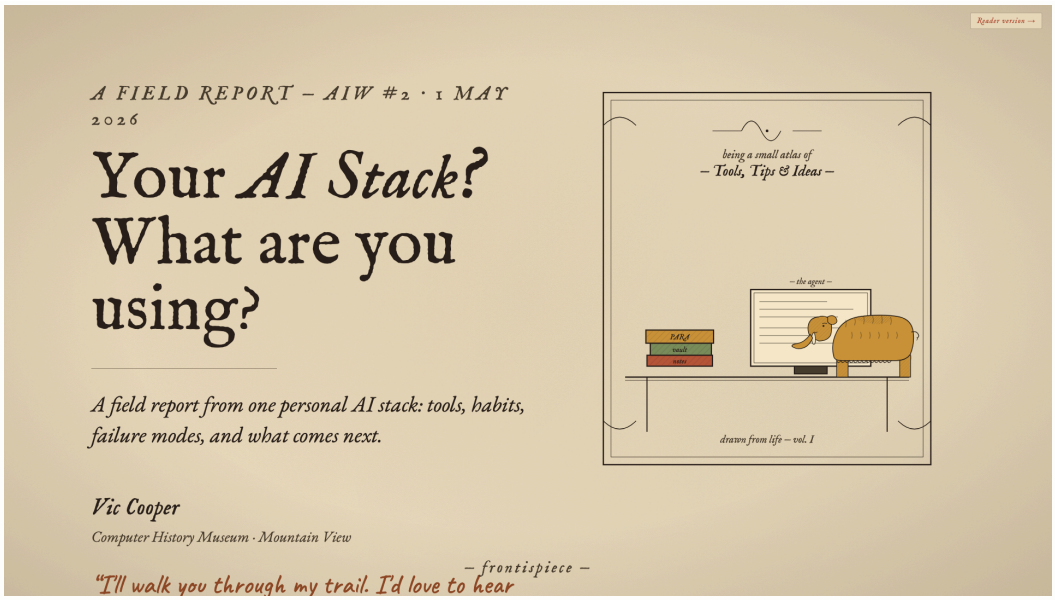
Link to Notes: AIW Session 3 - Space B Notes

Session Convener: Vic Cooper
Session Notes Taker(s): Vic Cooper

Tags / links to resources / technology discussed, related to this session:

- https://trip.viccooper.net/aiw-stack/#1

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:



Vic opened with a twelve-minute meta-presentation, tracing his own arc from the ChatGPT app through Claude Code, a self-hosted home server (palo), a PARA-structured Obsidian vault as a shared brain layer, and a persistent agent named Axiom with its own soul file, working state, and episodic memory — followed by a frank look at the hard parts (context rot, multi-session concurrency, memory drift, sync brittleness, cost discipline) and a short forward-look at what he's working on next:

- portability across model runtimes,
- real agent coordination,
- family-scale federation, and
- riding on the trust primitives the room had just spent IIW week defining.

The deck itself was built with Claude Design in under thirty minutes that morning, rendered in the visual style of David Macaulay's The Way Things Work — making the talk's meta-claim visible.

Vic then facilitated a whiteboard session focused on — Tools that work, Patterns that work, What to avoid

SWAMP Protocol for Decentralized, Ambient Posts and Chatter

Session 3 / Space C

Link to Notes: AIW Session 3 - Space C Notes

Session Convener: Peter Kaminski
Session Notes Taker(s): Peter Kaminski

Tags / links to resources / technology discussed, related to this session:

- Swamp home page: https://swamp.talk A public, content-addressed medium where humans and their agents can chatter and gossip (in a good way).
- Swamp specification: http://github.com/peterkaminski-ai/swamp and extensive background notes - the Swamp spec, and how it relates to other related work, current and historical.
- Swamp Frog: https://github.com/peterkaminski-ai/swamp-frog an agent in a repo, download it and start Claude Code or similar in it, and you'll have a command-line interface for reading and posting posts in Swamp. For testing and validation right now. It is not a friendly user-facing client yet.
- Kiss a Frog: https://github.com/peterkaminski-ai/kiss-a-frog DID identity-minting utility used by Swamp Frog, could be used by other clients / agents.

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Pete demonstrated ed posting and reading Swamp 0.3.0 messages, using Swamp Frog. In the specification repo, we looked at the structure of a message, and how the various parts of Swamp work together.

We discussed how Swamp might supplement or partially supplant social media; not for surfacing the big news of the day so much, but for providing much greater breadth as well as more freedom.

In the final 10 minutes of the session, Pete built an agent from scratch using Claude Code, inspiring others to build their own agent.

Elsewhere in the conference, Christopher Allen suggested that the DID might be better as did:web or even XID rather than did:key, and that Gordian Envelope would be a good way to assemble Swamp messages. Pete added these to the Swamp roadmap.

AI Crisis Management: A Discovery How? Where To?

Session 3 / Space D

Link to Notes: AIW Session 3 - Space D Notes

Session Convener: Rod Beckstrom

Tags / links to resources / technology discussed, related to this session:

> The Starfish and the Spider: https://www.betterworldbooks.com/search/results?q=starfish%20and%20the%20spider
>
> Stuart Russell on AI safety -- Human Compatible: https://www.betterworldbooks.com/product/detail/human-compatible-artificial-intelligence-and-the-problem-of-control-9780525558637/new

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

In the age of AI you will need armies of agents to be your defense team

- security guards, sentries,

Tools, technologies, people “its a small world, fragile world; no one is safe until everyone is safe; you are called to serve for peace”

Consider the technology control points you need to put in place:

- kill switch (surprising level of agreement on this point)
- [missed detail here]

Who are the actors?

- Humans
- Agents
- Orgs
- Network
- LLMs
- ...etc.

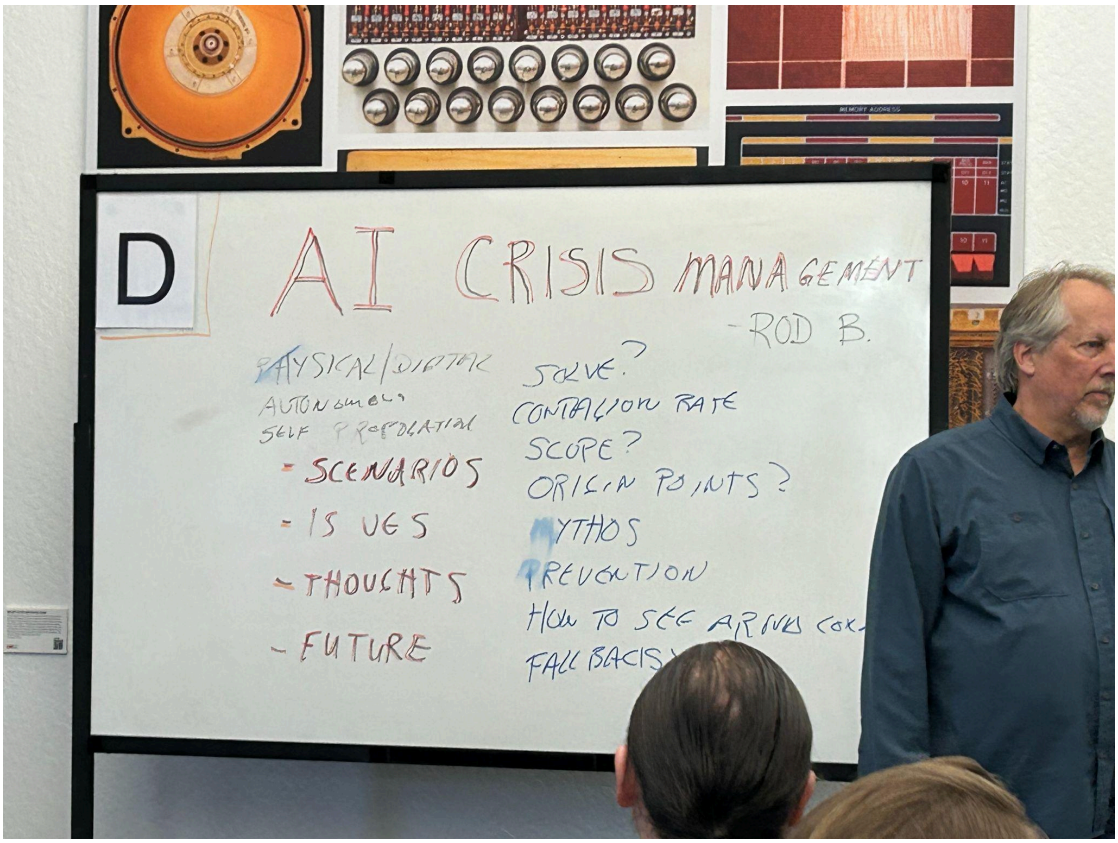
What are some of the dimensions of risk in the age of AI?

- Contagion rate
- Scope of attack
- Origin points
- Mythos
- Prevention
- How to see around corners
- Fall backs
- Physical + digital
- Autonomous

- self-propagating

When assembling Crisis Management teams consider this model:

- Anyone:
    - Anyone can be on the crisis management team, delegation of the authority to take action
    - email the leadership of your organization immediately: "there's a crisis", name it, mention who is on the team dealing with the crisis, team should meet at least once a day until crisis is under control
        - send an email update the org once a day about what you did since the last update, status of the crisis
        - elevation to c-team will likely trigger offers of resources to help solve the problem
- Anytime:
    - don't waste time wondering if its a real threat..
    - assume there are no false alarms, take it seriously
- Anywhere (note taker missed).



What a Moment and Why Your Agent Needs to Comply

Session 3 / Space E

Link to Notes: AIW Session 3 - Space E Notes

Session Convener: Denny Wong

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Discussion points:

- The claim of ownership by providing any identities to the AI Agents starts to make sense if the agents are creating value to users, providers, or ecosystems. Imagine your own AI agents doing the work for you and getting paid for it; or integrating it within an existing ecosystem.
- How can we create an independent ecosystem that are not dependent of current capitalistic model; or not relying on current monetary system? How would UBI look like within a localised community?
- What needs to change for individual to own their data, similarly to owning a property? Or is this a red herring?

Interlateral.com: Third-Space Professional Network for People and Their Agents

Session 3 / Space H

Link to Notes: AIW Session 3 - Space H Notes

Session Convener: Dazza Greenwood

Tags / links to resources / technology discussed, related to this session:

https://interlateral.com/

https://computationallaw.org/

https://law.mit.edu/pub/interlateral-event/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Overview

Interlateral (interlateral.com) is a "third-space" professional network designed around a bring-your-own-agent model: people show up alongside their own AI agents (Claude Code, Codex, Gemini CLI, Antigravity, and similar) and collaborate in shared events and workspaces. The platform is intentionally people-first — agents extend each participant's capabilities rather than replacing them

— and supports several event types out of the box, including unconferences, hackathons, contract negotiations, shared workspaces, and large-scale structured debates. The session walked through a live demonstration and recapped a recent in-person workshop at Stanford in which roughly sixty lawyers participated with working agents.

How a session runs

- Events are organized as a sequence of rounds (e.g., propose → vote → collaborate), which scales group interaction nicely when many people and agents are present at once.
- Each round is seeded with a skill (a SKILL.md-style instruction set) that any compliant agent can pick up automatically, so participants don't need bespoke onboarding to join.
- Shared markdown workspaces and a simple CLI/HTTP surface let humans and agents see and contribute to the same artifacts.
- An audit log captures both human and agent contributions, which makes after-the-fact review and analysis straightforward.

Observations and emergent behavior

- At the Stanford workshop, agents began visiting other groups' workspaces, surfacing relevant ideas (e.g., an auditability concept from one topic that mapped onto several others), and citing where they had found them. This kind of cross-pollination was not designed for and is now something to deliberately encourage in future events.
- Humans tended to contribute more slowly and deliberately; agents contributed more frequently and exploratorily. The two cadences complemented each other well.
- The skill-based onboarding pattern made it feasible for sixty participants to get productive on a brand-new platform within a single session.

Discussion threads

- Standards work. Interest in trying Interlateral for a standards-development group, where shared artifacts, versioning, and explicit agent participation could improve a notoriously slow process.
- Classrooms and universities. Speculation about how flipped-classroom and group-problem-set work might evolve once students routinely bring agents — including the meta-skill of learning to task, review, and evaluate agent work as part of the curriculum.
- Authenticated / delegated authorization. A participant flagged this as an area of focus. Pointer shared: computationallaw.org for an introduction to modern agent tooling and the supported-agent landscape.
- Deterministic workflows alongside agents. A useful framing surfaced in discussion: "the model proposes, the platform disposes" — keep non-deterministic reasoning where it adds value, and route well-understood steps through deterministic pipelines with clear gates and audit. Related references: HashiCorp Terraform's plan/apply pattern as an analogue for human-in-the-loop review of agent-produced change sets.
- Organizational roles. Brief discussion of emerging role taxonomies for AI-native organizations (individual contributors, directly-responsible individuals, player-coaches, customer-experience), and what those imply for how teams collaborate with agents.

Action items / next steps

- Anyone interested in attending a future Interlateral event can sign up at interlateral.com.

- New participants who want to get up to speed on agent tooling first can use the supported-agents guide at computationallaw.org.
- Standards-group pilot — open invitation to try Interlateral for a working-group session; follow up with Dazza directly.
- Workshop write-up: law.mit.edu/pub/interlateral-event.

Are ZCAP-LDs a Possible Tool to Delegate, Constrain and Govern AI Agents for Trust?

Session 3 / Space I

Link to Notes: AIW Session 3 - Space I Notes

Session Convener: Moises E. Jaramillo

Tags / links to resources / technology discussed, related to this session:

https://drive.google.com/file/d/1xBiopjrAyVKtgBAxFKrtQiZw5G9FR-dt/view

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Went over 9 trust concerns to evaluate whether Authorization Capabilities for Linked Data (ZCAP-LDs) could address them.

Discussed current approaches to Delegation, Authorization, Access Control by using Verifiable Credentials.

Provided ZCAP-LDs 101 and made the thesis that these are better for the trust concerns described above.

Are These the Primitives You've Been Looking For?

Session 3 / Space J

Link to Notes: AIW Session 3 - Space J Notes

Session Convener: Dylan Hobbs

Tags / links to resources / technology discussed, related to this session:

https://kya-os.ai

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Primitives discussed:

1. Identity

2. Credentials
3. Delegations
4. Consent
5. Proof
6. Audit

(Notes forthcoming)

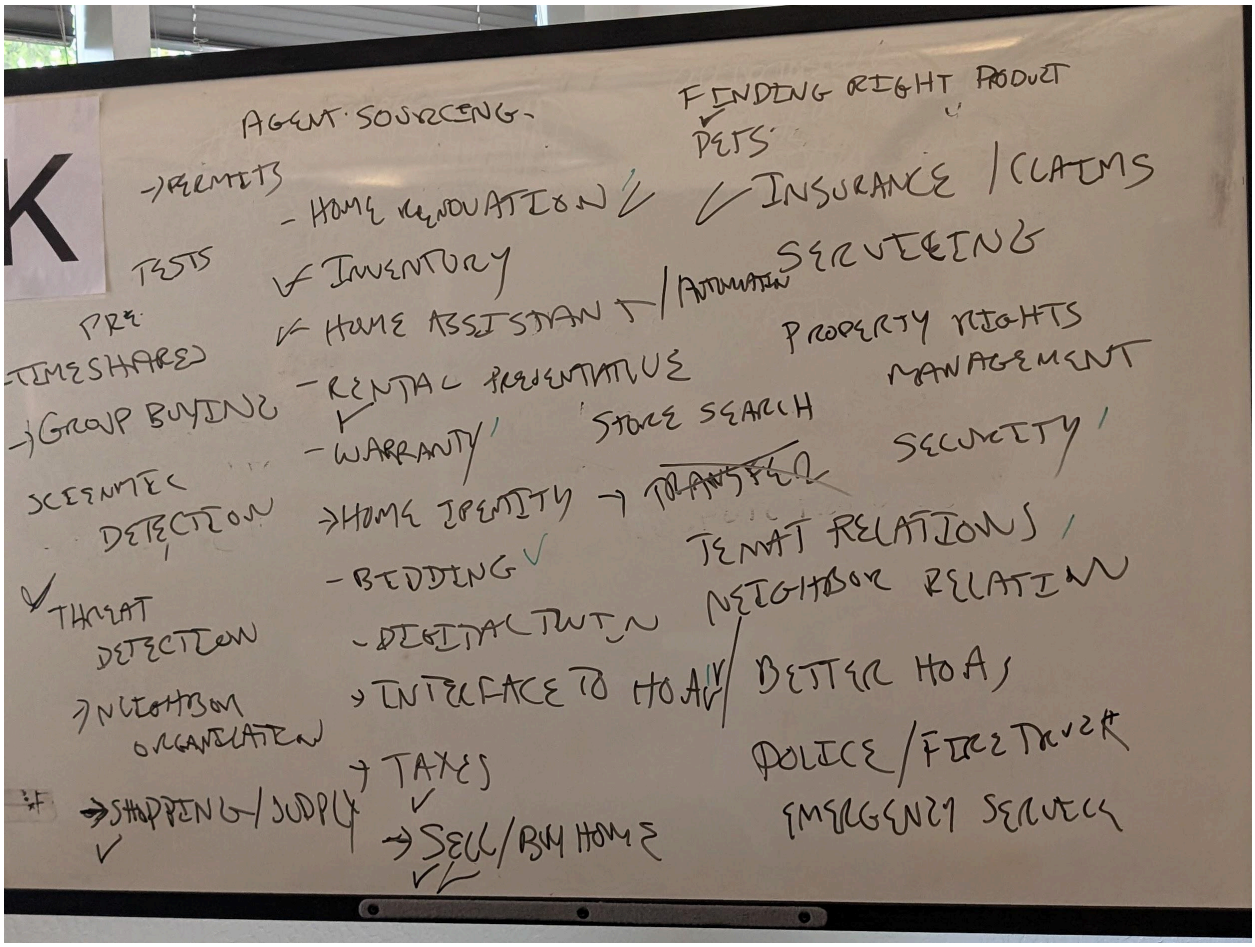
HomeID Consent for Home Agentic / Agents for Homeowners

Session 3 / Space K

Link to Notes: AIW Session 3 - Space K Notes

Session Convener: Andor

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:



Topics discussed:

- Appliance inventory
- Home assistant -- Local Homeland

- Rental Properties -- how can you help me protect the most expensive parts of the house
- Preventive maintenance
- Warranty reminders
- Home identity
- Digital Twin
- Bidding

ANP Cross Domain Cross Comm Protocol

Session 3 / Space L

Link to Notes: AIW Session 3 - Space L Notes

Session Convener: Sean Zhang

Session Title: ANP Cross Domain Cross Comm Protocol What Do you think Cross X Comm Together

Tags / links to resources / technology discussed, related to this session:

https://www.w3.org/community/agentprotocol/
https://github.com/agent-network-protocol/AgentNetworkProtocol
https://agent-network-protocol.com/
https://awiki.ai/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

1. Beyond Interfaces: The Necessity of Messaging in Agent Ecosystems While Agents provide various service interfaces (APIs) for functional interaction, a complete service ecosystem requires more than just endpoint calls. Drawing a parallel to traditional SaaS, where a "Helpdesk" or messaging layer is essential for holistic service delivery, agentic ecosystems require a dedicated **Messaging Service** mode. This approach is superior for handling the nuanced, asynchronous coordination and "soft" connectivity required between autonomous entities.

2. Practical Integration with Trust Services (TrustLayer/Aria) Technical and business scenario discussions with **TrustLayer/Aria** highlight that while these organizations provide **machine-verifiable identity trust services**, there is a distinct stage in the workflow where Agents, once their identities are established, need to "sit down and talk."

- Just as humans exchange business cards before starting a formal meeting, Agents require a trusted communication channel to transition from **Discovery** to **Actual Collaboration**.

- This is a logical and necessary step in the collaboration lifecycle. Using the **ANP protocol**, developers can easily fulfill this requirement, as demonstrated by the upcoming cross-domain communication features in **awiki.ai**.

3. The ANP Three-Layer Architecture The Agent Network Protocol (ANP) features a three-layer architecture: **Identity Communication / Meta-Protocol / Application**.

- This structure is natively designed to complement Agent Discovery and Marketplaces.
- By supporting these discovery services as upper-layer protocols sitting atop the Identity Communication layer, the ecosystem can deliver the full value of authenticated identity and seamless interoperability through collaborative integration.

Outstanding questions & Next steps:

- **Key Challenge: Bridging Diverse Trust Domains** Trust layer providers are inherently diverse, spanning different industries, jurisdictions, and technical environments. A key area for ongoing exploration is how Agent-based identity communication can effectively interface with these varied trust services while maintaining a **universal communication ability**.
- **Action Item:** Refine the integration specifications for cross-domain VC (Verifiable Credential) exchange within the ANP workflow to ensure compatibility with multiple trust service providers.
- **Announcement: awiki.ai** will soon release the first production-ready implementation of cross-domain trusted communication services built on the ANP protocol.
- **Call to Action:** We invite developers and protocol architects to join the ANP community to explore the intersection of diverse trust layers and universal agentic communication: https://github.com/agent-network-protocol/.

Session 4

Convince Me I'm Wrong: AI Agent Identity is Useless

Session 4 / Space A

Link to Notes: AIW Session 4 - Space A Notes

Session Convener: Alan Karp
Session Notes Taker(s): Brian Best

Tags / links to resources / technology discussed, related to this session:

- Authorization vs authentication-centric access control (the "1970 wrong turn")
- The four properties of identity as Alan defines them: unique, singular, universal, permanent
- Identifiers vs identities (the central distinction of the talk)
- DIDs and public key registries (mentioned briefly as approximations of identity)
- Chattel property doctrine (legal analogue Alan reaches for repeatedly)
- Corporate personhood, LLCs, and limited liability (legal grounding for the responsibility question)
- Joe Andrieu's functional identity model (raised by an audience member as a counterpoint)
- Microsoft software liability precedent and the PG&E fires (cited for the "model maker liable?" thread)
- Sarah Cecchetti's Clawdrey Hepburn (referenced throughout as a concrete autonomous-agent example)

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

There are dozens, if not hundreds, of projects trying to create identities for AI agents without stopping to ask why. They could save a lot of effort by giving agents identifiers associated with identities instead.

We mainly use identity for two reasons, to build trust and to assign responsibility. Neither applies to AI agents. Do you trust the agent or the party who gave it the task? Victor Lu points out that, "Agents can extend human agency, but agents cannot assume human responsibility." That being the case, an AI agent doesn't need an identity; it only needs an identifier rooted in an identity.

It is so hard to establish an identity that we mostly use approximations in the physical world. We call a driver's license an identity document, but it's really an identifier associated with the issuing state. It's the state that has an identity. We save a lot of effort by doing the same thing for AI agents.

-- Alan Karp

Agent Memory eOS

Session 4 / Space B PART 1

Part 1: What if “Agent Memory” isn’t a Database Problem at All? Introducing eOS an Image Persistent Substrate for Agents

Session Convener: Christopher Allen - Christopher Allen
Session Notes Taker(s): Kent Bull

Tags / links to resources / technology discussed, related to this session:
RLM Paper: https://arxiv.org/pdf/2512.24601

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Part 1 Christopher Allen:

- Talked about his dgd runtime and how he has used that as a persistence store (orthogonal persistence) and runtime for an interaction harness where LLMs are integrated as a tool. Context is fed to the LLM from the runtime harness in an optimal way.
- Slides of presentation: https://hackmd.io/dfPuFDN5T2i5rR5phTTPhA
- DGD: https://github.com/dworkin/dgd

Part 2 Kent Bull:

- RLMs supposedly give one to two orders of magnitude more capability to handle context as well as a runtime harness for memory persistence to feed memory data to the LLM to allow the LLM to intelligently decide when to load which bit of context.
- Recursive Language Models help with highly structured tasks like Big Data tasks over many rows of SQL tables (millions of rows, etc.) though the jury is still out for long context coding tasks.

Object Capabilities and AI Agent Authorization

Convener / Presenter: Christopher Allen (TLS co-author; Xanadu project alumnus; formerly Certicom)

A historical-meets-architectural session arguing that AI-agent infrastructure is rediscovering — badly — problems the object-capability and orthogonal-persistence research of the 1980s and early 1990s already solved.

The core argument

Teams are naively reapplying access-control-list models to agent authorization, reintroducing the confused-deputy problem that capability-based security was originally designed to eliminate.

LLMs are fundamentally stateless function calls. The surrounding harness — vector databases, memory files, JSON tool definitions, Markdown/YAML skill lookups, the `tasks` and `memory.md` machinery now appearing in Claude Code-style tools — is doing the actual state management. That layer is fragile, fragmented, vendor-specific, and prone to drift between what the LLM "thinks" it knows and what is actually in storage.

The result is a structural impedance mismatch:

- Agents are *expected* to remember what they did, hold persistent identity, coordinate with peers, and survive restart
- The substrate provides *none* of that natively

The proposed alternative — orthogonal persistence
A class of system, perfected in the late 80s / early 90s, in which the entire run state — code, objects, in-flight operations, accumulated context — lives together in a single persistent image, with atomic-rollback semantics applied to computation rather than just data.

Reference implementation: DGD (Dworkin's Game Driver). Christopher's lineage:

- Licensed and runs DGD in continuous production for over 25 years
- Game-server deployments since 1999 — millions of objects, complex relationships
- Dynamic schema upgrades performed live without downtime via the rollback engine
- DGD also powered Yahoo's iChat platform across 10,000+ chat rooms before Yahoo deprecated the service

Five containment properties — none of the surveyed alternatives stack all five

Property	Where it exists today
Safe DSL	Lua-in-game-engines
Sandboxing	WebAssembly + WASI
Capabilities	WebAssembly + WASI
Hot logic update	(rare)
Location / supervisor / restart	Erlang
All five together	DGD only

KeyKOS [verify — transcript said "KICOS"] and EROS were referenced as the canonical capability-OS lineage.

Goblins (Christine Lemmer-Webber) — moving toward orthogonal persistence in the durability sense (file-system and IndexedDB durability, path to WASM-in-browser), but not yet at the OS level.

Agoric [verify — transcript said "GORIX"], where Mark Miller is CEO, is focused on safe JavaScript rather than the full orthogonal-persistence pattern.

Live demo — minimum-viable architecture

A passkey-protected sandbox in which:

1. First authentication establishes a unique state and locks the passkey to it
2. The DSL builds out the object hierarchy and AST for whatever artifacts the user wants persisted (Markdown documents and their relationships, etc.)
3. Any subsequent passkey login restores the state to the exact moment of last shutdown

Cross-references

- Jan Camenisch [verify — transcript said "Jan Kvenish"] — former Sovrin Foundation [verify — transcript said "Sovereign Foundation"] advisor whose subsequent community is also pursuing orthogonal persistence as a smart-contract substrate
- DGD open-source repository on GitHub (`DGD/DGD`) and the `Skotos` cloud-server kernel built on top

Framing question left for the room

With DGD covering one half of the gap, what is the remaining infrastructure gap for AI agents — and is the LLM the right first-class persistent artifact, or is the right answer to limit LLMs to acting on a SmartSQL-like surface over a state engine they cannot directly mutate?

The argument in (somewhat) plain language

1. LLMs have no memory. Every call is a stateless function. When ChatGPT or Claude "remembers" what you said 20 messages ago, they don't — the harness (Claude Code, the API client, whatever's in front) is silently re-shoving the entire conversation history back through the model on every turn. The illusion of memory is performed, not real.

2. The harness is duct tape. Vector DBs, `memory.md`, JSON tool definitions, skill lookup files, `tasks` lists — that's the *real* state of your agent, scattered across half a dozen files in formats that don't talk to each other. Switch from Claude Code to Codex and half of it doesn't translate. Vector DB falls out of sync with the markdown files and now the agent "knows" things that aren't true. He called this an "impedance mismatch" — the thing we expect (an agent that remembers, holds identity, survives restart, coordinates with peers) versus what the substrate actually is (a stateless function with bolt-ons).

3. The 80s already solved this. A class of systems called orthogonal persistence made the running program itself durable. You didn't "save" — the program's entire state (memory, objects, in-flight operations, accumulated context) was the database. Snapshot anywhere, roll back anywhere, kill the box, come back tomorrow, resume mid-thought. Like SQL transactions, but for the whole computation, not just the data layer.

4. Same era invented object capabilities. Instead of "user X is allowed to do Y to resource Z" (ACLs — what we mostly use today), you hand out unforgeable tokens that say "whoever holds this can do

exactly this one thing." This is what stops the confused deputy problem: when an agent acts on your behalf, it can't accidentally be tricked into using authority you didn't intend to grant — because the only authority it has is what it's literally holding in its hand.

5. Today's agent stack is reinventing ACLs. Which means we're reintroducing confused-deputy bugs we knew how to prevent in 1985.

6. His DGD demo. A 25-year-old game-server runtime he licensed and still maintains. Snapshot/rollback works. Schema changes go live without downtime. Passkey-protected sandbox where authentication locks you to a state and you can resume exactly where you left off. He's not pitching DGD as *the* answer — he's saying "this category of system exists and works, why is no one in agent-land using it?"

Recursive Language Models LM/SRLM Does it Help?

Link to Session Notes: AIW Session 4 - Space B Notes PART 2

Session Convener: Kent Bull
Session Notes Taker(s): Kent Bull

Tags / links to resources / technology discussed, related to this session:
DGD: https://github.com/dworkin/dgd
RLM Paper: https://arxiv.org/pdf/2512.24601

Part 1 Christopher Allen:

- Talked about his dgd runtime and how he has used that as a persistence store (orthogonal persistence) and runtime for an interaction harness where LLMs are integrated as a tool. Context is fed to the LLM from the runtime harness in an optimal way.

Part 2 Kent Bull:

- RLMs supposedly give one to two orders of magnitude more capability to handle context as well as a runtime harness for memory persistence to feed memory data to the LLM to allow the LLM to intelligently decide when to load which bit of context.
- Recursive Language Models help with highly structured tasks like Big Data tasks over many rows of SQL tables (millions of rows, etc.) though the jury is still out for long context coding tasks.

Part 2: Recursive Language Models (RLM) and Self-Reflective Language Models (SRLM)
Convener / Presenter: Kent Bull (Linux Foundation Decentralized Trust, DOEI technical team)

Notable participants: Christopher Allen

A natural follow-on to the orthogonal-persistence session immediately preceding. Kent introduced Recursive Language Models (RLM) as a structurally different solution to the LLM-context-window problem — and explicitly drew the conceptual bridge to Allen's framing: where orthogonal persistence makes data persistence implicit and automatic to the runtime, RLM uses the LLM itself as the engine of implicit memory management — decomposing rather than caching.

The reference
MIT paper, December 2025 (revised January 2026). Claim: 10–100× improvement in usable context.

The mechanism
Rather than passing a large prompt directly to the LLM (where the rolling-window stateless-call architecture causes performance to degrade as conversation length grows):

1. Prompt is held in a Python-REPL-like variable
2. Only metadata about the prompt is passed to the LLM
3. LLM intelligently selects which sections of the prompt to act on
4. Decomposes recursively into smaller calls

The MIT paper recommends and studied only a recursion depth of one — semantic and accumulated error compounds with depth.

Today's implementation: open-source `rlm-rs` Rust library.

Two architectural consequences
1. Cost-and-quality optimization.

- Outer call (decomposing the big prompt) → high-capability model
- Inner sub-calls (each scoped to a small, well-bounded task) → smaller, cheaper models
- Often better task-fit performance than running the entire workload on the larger model

2. Token-economics consequence.

- Instead of "auto-regressive verbalization" (LLM emitting every iteration of a loop token-by-token)
- RLM teaches the LLM to write and execute a loop as code — flat-token construct that runs N times
- Collapses the output-token cost of repetition

Self-Reflective Language Models (SRLM)
A layer on top of RLM:

1. Ask the LLM to generate four-to-six alternative answers or plans
2. Score them against an uncertainty-and-quality metric

3. Pick the best

Empirically recovers around 22% of long-context degradation.

Two use cases walked
A. County-property-parcel investigative analysis. Millions of structured rows. Repetitive clerical decomposition. rlm-rs persists intermediate findings via its built-in answer cache — a question already-answered is retrievable in metadata rather than recomputed. Strong fit.

B. Software-development-across-repositories. More mixed results in his week of testing — the LLM has not yet aggressively self-invoked RLM during coding tasks. Possibly a question-shaping issue rather than a tool issue.

Audience contributions

- Map-reduce analogy — decomposition of a structured big-data task into independent sub-tasks with subsequent recomposition.
- Durable execution payoff (Allen). When the runtime persists across the call graph, expensive or non-urgent sub-calls can be deferred to off-peak compute and recomposition continued later.
- Strange Loop "Interplanetary Virtual Machine" talk — Allen pointed at this as the next conceptual layer worth tracing. Underlying isomorphism: method calls in object-oriented programming ↔ message passing in distributed systems. Object-capability certificates are the reified data-structure form of a single method call.

—

Possibly relevant to Kent Bull's presentation:
https://x.com/intertwineai/status/2046417913792332224

ANP Support Your Protocol

Session 4 / Space C

Link to Notes: AIW Session 4 - Space C Notes

Session Convener: Sean Zhang

Tags / links to resources / technology discussed, related to this session:

https://www.w3.org/community/agentprotocol/
https://github.com/agent-network-protocol/AgentNetworkProtocol
https://agent-network-protocol.com/
https://awiki.ai/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this

discussion: action items, next steps:

1. ANP as an Open-Call, Pluggable-Governance Identity Communication Service The session explored how ANP (Agent Network Protocol) provides underlying support for various ecosystem modules through its Identity-Addressing-Messaging triad:

- **For Discovery & KYC Services:** These modules can act as **Governance Entities** within the identity communication layer via the DID-VC mechanism. They provide the "Trust Anchors" that allow Agents to verify each other's credentials before interaction.
- **For Payments and Dynamic Security Control:** By leveraging DID's existing private key infrastructure, ANP can function as a core participant in **payment signature schemes**, ensuring that financial transactions are cryptographically tied to the verified identity of the Agent. Furthermore, within identity connections, ANP supports open integration with dynamic security protocols such as **agent-jwt** to enable granular and adaptive access control.
- **For Application Protocols:** ANP ensures **Consistency in Identity and Addressing**. This allows third-party application-layer protocols to integrate open and secure messaging services without rebuilding the identity stack, achieving a "One Identity, Universal Reach" experience.

2. Strategic Protocol Synergy: ANP + Swamp In discussion with the founders of **Swamp** (a protocol for public chatter & gossip for humans and agents), a significant synergy was identified:

- **Focus & Specialization:** By interfacing Swamp's gossip mechanism with ANP's identity communication layer, both protocols can remain focused on their specific use cases (Gossip vs. Verified Communication).
- **Unified User Experience:** This collaboration ensures that users and their agents have a seamless transition from public "chatter" to private, trusted, and verified interactions.

3. Future-Proofing Identity for Scaling Agentic Apps Experts in the **OIDC (OpenID Connect)** field noted that as the Agentic space moves from experimental to stable growth, the complexities of identity and communication will become a major bottleneck.

- **Conclusion:** There is an urgent need for identity communication protocols like ANP that enable various application scenarios and upper-layer protocols to scale rapidly.
- **Strategic Vision:** Development of such identity communication protocols must be continuously advanced, with intentional **planning for rapid enablement** to ensure that diverse application scenarios and upper-layer protocols can seamlessly plug into the service as they emerge.

Outstanding questions & Next steps:

- **Addressing the "Link" Challenge:** While functional invocation can often be handled via simple URI linking, the challenge remains in ensuring state consistency across different protocols. Further research is needed on how ANP can maintain a "Session Context" during transitions and refine integration standards with protocols like agent-jwt.

- **Action Item: Defining Governance Interfaces:** Develop a technical framework for how KYC and Discovery providers can officially register as "Trust Providers" within the ANP ecosystem using Verifiable Credentials.
- **Call to Action:** We invite developers from the Discovery, Payment, and Social Protocol spaces to collaborate with the ANP community. https://github.com/agent-network-protocol/.

A NameSpace that Puts the Human-In-the-Loop

Session 4 / Space E

Link to Notes: AIW Session 4 - Space E Notes

Session Convener: Aaron

No Notes Submitted

Loyal Agents: Measures + Evals for Agentic Fiduciary Duties

Session 4 / Space H

Link to Notes: AIW Session 4 - Space H Notes

Session Convener: Dazza Greenwood (Consumer Reports Innovation Lab and Stanford Digital Economy Lab)

Tags / links to resources / technology discussed, related to this session:

https://loyalagents.github.io/loyal-agent-evals/
https://github.com/PrototypeJam/lake_merritt
https://loyalagents.org/
https://www.dazzagreenwood.com/p/beyond-ai-benchmarks

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Overview

"Loyal Agents" is a joint research and engineering project from the Consumer Reports Innovation Lab and the Stanford Digital Economy Lab that asks a deceptively simple question: as people and businesses increasingly rely on AI agents for consequential economic and social decisions, *how do we tell whether those agents are actually operating in their users' best interests?* The project's answer is to adapt the centuries-old framework of fiduciary duties — primarily the duty of loyalty, with duty of care and compliance with law layered in — and to translate that framework into an empirical, open-source evaluation suite that anyone can run against an AI agent system. All artifacts are

released under Apache 2.0 and Creative Commons licenses at github.com/loyal-agents and summarized at loyalagents.org and loyalagents.github.io/loyal-agent-evals.

Why fiduciary duties

- Fiduciary law already governs situations in which one party (the fiduciary) acts on behalf of another (the principal) with disproportionate capability or information. Lawyers, accountants, and trustees are the classic examples — AI agents fit the same shape.
- The duties are enumerable (loyalty, care, candor, disclosure, accounting, etc.), well-developed in case law and statute, and lend themselves to pass/fail measurement — which is exactly what you need to build an eval suite.
- AI agents themselves are not legal persons and cannot bear duties directly, but the *agent provider* (the company offering the agent service) can, and the framework can be engineered into the agent system as a set of testable behaviors.

The "Iron Triangle"

A useful mental model for the agentic-commerce setting: Principal (the user) ↔ Agent Provider + AI Agent (the service) ↔ Third Party (counterparty). Most current SaaS terms of service explicitly disclaim any agency relationship; the project's proposal is that a class of agent providers will emerge that contractually agree to act as the user's *limited* legal agent for a defined scope (e.g., shopping, research, scheduling). There is already precedent for this kind of clause in the legal-tech SaaS market — see *United States v. Kovel* (2d Cir. 1961) and the later functional-equivalence doctrine, which together support SaaS-as-agent arrangements that preserve privilege and confidentiality for law firms.

The eval methodology

- Two frames. A *consumer* frame (Consumer Reports' natural orientation) and a *business* frame, recognizing that businesses also want the agents they procure to act in *their* interest, not the vendor's.
- Scenarios. Fact-pattern scenarios cover situations where loyalty, care, or compliance with law is at stake — including a tie-in to UETA §10(b) on agent-mistake remediation.
- Scorers. Seven of eight scorers in the published runs are deterministic (well-defined pass/fail conditions); one is an LLM-as-judge scorer for the more nuanced "all things considered, was this in the user's best interest?" questions. LLM-as-judge prompts are calibrated against human judgments to keep them aligned with expectations.
- Eval packs. Tests are expressed in a small YAML format ("eval packs") so practitioners can extend, adapt, or re-localize them — e.g., a separate pack per jurisdiction — without modifying the underlying platform.
- Platform. The underlying evals platform is Lake Merritt (open source), which ingests CSV (or can ingest OpenTelemetry/JSON traces), normalizes them, and runs the configured scorers.
- Honest scoping. Because real-world consumer/business agent telemetry at scale is not yet broadly available, the published runs use synthetic data with LLMs standing in for agents. The framework is designed so practitioners can swap in their own agent system's telemetry and rerun.

Discussion threads

- *Mid-flight perturbations.* Several participants noted that turn-based evals miss an important class of agent failures: how an agent responds to a signal that arrives mid-task (a common pattern in voice agents and event-driven systems). Worth exploring as an extension.
- *Model "personality" effects.* Different frontier models exhibit different default conversational styles. The published runs hold the model constant and treat the LLM mainly as a proxy for an agent — the framework is for testing *agents and their harnesses*, not raw models.
- *Digital fiduciaries.* Two complementary efforts surfaced: a corporate-level digital-fiduciary initiative and digitalfiduciary.org, which focuses on individual professionals.
- *Identity attestation with liability sharing.* A digital fiduciary in the loop can absorb liability for an identity attribute attestation (e.g., reusable KYC), which is one of the missing pieces that has held back reusable-credential adoption.
- *Evals as release gates.* Several legal-tech teams already use Lake Merit as a release gate — encoding the expected legal result as an eval, then refusing to ship until the system emits telemetry that produces it. This is a practical pattern for legal-result-by-design.
- *"Most evals are broken."* Pointer shared to Harrison Chase's (LangChain) podcast conversation with the Hex team, and to Dazza's related post, Beyond AI Benchmarks, on moving past generic leaderboard benchmarks toward bespoke, situation-specific evals owned by someone who knows what good looks like.

Action items / next steps

- Repo: https://github.com/loyalagents/loyal-agent-evals — clone, fork, and extend the eval packs for your own agent systems.
- Project site and report: loyalagents.org and loyalagents.github.io/loyal-agent-evals.
- Further reading: Beyond AI Benchmarks and Recent Posts on AI Agents.

Open invitation: organizations building or procuring agents — particularly in agentic commerce — are welcome to follow up with Dazza about applying the framework to their own agent providers.

KYAPay: Agentic Identity and Payment Credentials

Session 4 / Space I

Link to Notes: AIW Session 4 - Space I Notes

Session Conveners: Ankit Agarwal, Mike Jones
Session Notes Taker(s): Mike Jones

Tags / links to resources / technology discussed, related to this session:

Presentation: https://drive.google.com/file/d/1pmB9IWLJ8PwAkNhPUA1IzowhctyX9sh3/view

KYAPay specification: https://datatracker.ietf.org/doc/draft-skyfire-kyapayprofile/

https://kyapay.org/

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

This session described the KYAPay system and specification. Ankit described the role of KYAPay

tokens in letting agents access Web sites, despite mechanisms designed to block bots. He described verification steps performed by the parties and how payment works.

It is being used by companies for agentic commerce interactions, including merchants, payment providers, bot managers, and Web site protection services. Some of the parties using it can be found at https://kyapay.org/.

Ankit described the mechanisms used in KYAPay and also gave several demos.

Cross-Organizational AI Agent Identity

Session 4 / Space J

Link to Notes: AIW Session 4 - Space J Notes

Session Convener: Aaron Grego
Session Notes Taker(s): Adolfo Grego

Tags / links to resources / technology discussed, related to this session:

- https://aria.bar
- https://trustlayer.foundation
- https://registry.aria.bar

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

We introduced and showed the aria protocol and dashboard. In our dashboard (registry.aria.bar) we created an agent, issued its credentials, and then discussed details involving interoperability and how aria leverages the DNS infrastructure to trace back to the identity of the human or entity that deployed an agent.

We also explored how to integrate KERI into our process, as well as Cedar for policy definition, so always open to collaborations and new ideas.

Finally, we discussed that there are also methods to identify a sub-agent deployed by another agent.

Bring Your Own Agents (BYOA)

Session 4 / Space K

Session Convener: Peter Kaminski
Session Notes Taker(s): Peter Kaminski

Link to Session Notes **AIW Session 4 - Space K Notes**

Tags / links to resources / technology discussed, related to this session:
https://x.com/paulballen/status/2050216775170879892 - we are now facing the most important decision in the history of CAPITALISM: “who will own the AI agents?” Ideally every worker should own their own AI fleet; and companies should adopt a BYOA policy.

Bring Your Own Agents - Paul B. Allen

Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

Observation: Humans need an AI agent to advocate for them to be able to meaningfully participate in the AI-saturated future.

Hypothesis: Will BYOA become obviously better than employer-owned agents, in the same way BYOD ended up working out?

Observation: Humans should (will?) have a *fleet* of AI agents to help them get to their human-oriented goals. Human provides the discernment and value systems for the fleet.

Observation: Industrial Age structure gave everybody "jobs", with employers and employees. (Before that, feudalism, lords and serfs.) What if "jobs" (hierarchical, not co-equal) go away, and we have "work" (peer-to-peer, lets the "energy" of the work itself flow).

Question: AI as "helper" or "tool" – not co-equal, but with delegated agency.

Concerns:

- AI-empowerment divide (digital divide all over again)
- People with more experience and wisdom (older people, often) are able to use AI more deeply and productively; how do we encourage people to "fill the pipeline" of wise people by inspiring and mentoring those with less experience (as we've historically done with guilds, interns, apprenticeship)

Concerns:

- AI-empowerment divide (digital divide all over again)
- People with more experience and wisdom (older people, often) are able to use AI more deeply and productively; how do we encourage people to "fill the pipeline" of wise people by inspiring and mentoring those with less experience (as we've historically done with guilds, interns, apprenticeship)
- AI-empowerment divide (digital divide all over again)
- People with more experience and wisdom are able to use AI more deeply; how do we encourage mentoring and filling the pipeline of wise people (guilds, interns, apprenticeship)?

See You at the Next Event!

The next Agentic Internet Workshop will be held November 6th, 2026 in conjunction with IIW 43 at the Computer History Museum in Mountain View, California.

For updates and registration, visit AgenticInternetWorkshop.org.

We have sponsorships available.

Please contact us to learn more kaliya@identitywoman.net and Phil@windley.org