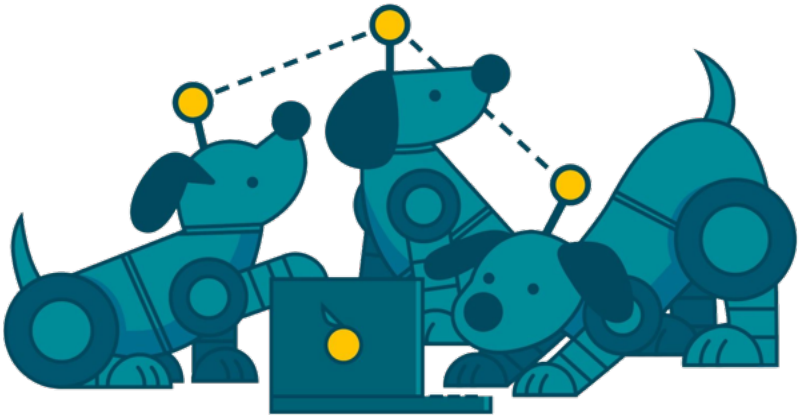
The Agentic Internet Workshop #1

October 24th, 2025

Computer History Museum, Mountain View, CA



Book of Proceedings

AgenticInternetWorkshop.org

Thank you to our sponsors:

[Consumer Reports]

[AWS Identity]

[JLINC]

[Glide Identity]

[AgentOverLay]

Founded by [Andor Kesselman], [Kaliya Young] and [Phil Windley].

Co-produced by Heidi Nobantu Saul & Kimberly Culclager

Contents:

About the AIW

Agentic Internet Workshop builds on the 20-year legacy of the **Internet Identity Workshop**, hosted in Mountain View, California. We are advancing the next generation of protocols that will define how AI agents connect, collaborate, and preserve human judgment in an increasingly agentic world.

Our mission is to provide a neutral forum for protocol definition and multi-stakeholder collaboration, with a vision to protect human integrity, judgment, and creativity as agentic systems become more prevalent.

The workshop was conceived by Kaliya Young (Identity Woman) and Andor Kesselman, and produced in collaboration with the Internet Identity Workshop Foundation and its Executive Director, Phil Windley.

Following the proven Internet Identity Workshop model, AIW employs Open Space Technology—an unconference format where participants collectively create the agenda during the opening circle. Any attendee can propose topics for discussion, ensuring the conversations reflect the community's most pressing priorities.

Agentic Internet Workshop Schedule

FRIDAY, October 24 / Doors Open at 8:00 AM for Registration	
Barista! And Continental Breakfast	8:00 - 9:00
Opening Circle / Agenda Creation	9:00 -10:00
Session 1	10:00 - 11:00
Session 2	11:00 - 12:00
Lunch	12:00 - 1:00
Session 3	1:00-2:00
Session 4	2:00-3:00
Closing Circle	3:00-4:00

Social Media about AIW

Matthias Moeller · 2nd
Strategic Technology Leader | Expert in AI, Blockchain, Inno...
1mo ·

+ Follow ···

The agentic web is taking shape — and trust is at the center of it.

🚢 Reflections from the **Agentic Internet Workshop** in Mountain View

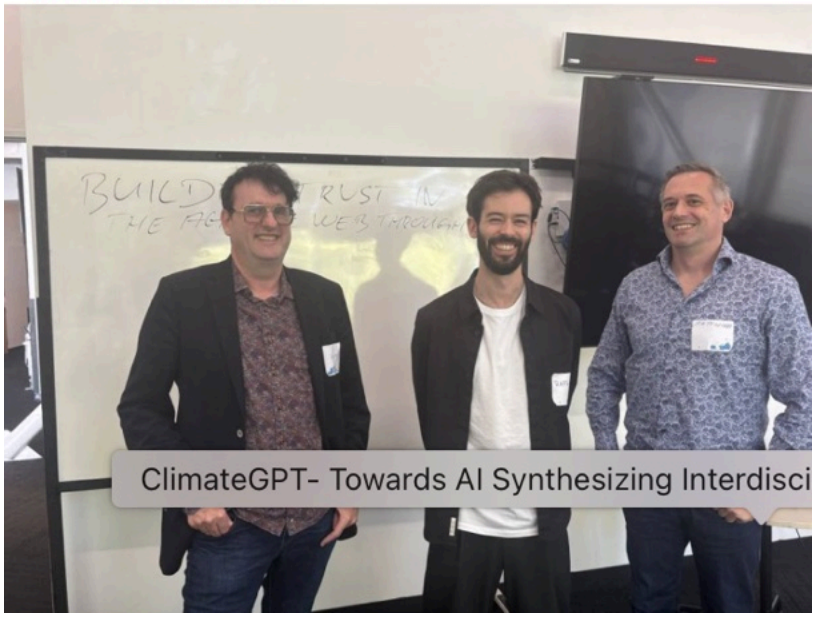
Last week, I participated in the Agentic Internet Workshop (AIW) — an **Internet Identity Workshop** -inspired event. Like IIW, it followed the Open Space Technology format: no pre-set agenda, just a room full of other AI enthusiasts, whiteboards, and discussions that evolve organically.

What made AIW special to me was its focus on building the foundations of the agentic web - not the next productivity hack, but the shared standards and protocols that will make the next phase of the internet's evolution (or revolution?) interoperable and trustworthy.

3andAI hosted a session on "How to Build Trust in the Agentic Web through Accountability."
Our main takeaway: the agentic web needs a trust layer that combines reputation with accountability. We introduced a collateral-based accountability mechanism, which received strong support and sparked meaningful discussion on how to embed trust into the fabric of agentic systems.

On a personal note, I really appreciate the Open Space format — it fosters genuine collaboration, open discussion, and a sense of shared ownership over the topics.

The agentic web is still in its early days — but it's moving fast — the next leap of the internet, adding "act" to read, write, and own.
#AgenticWeb #AgenticAI #AI #MCP #ERC8004 #AIStandards #Accountability #Trust #InternetIdentityWorkshop #AgenticInternetWorkshop



The First Agentic Internet Workshop

https://www.technometria.com/p/the-first-agentic-internet-workshop

Phil Windley Nov 06, 2025

Summary: The first Agentic Internet Workshop (AIW1) took place on October 24, 2025, the day after IIW 41, bringing together a global group to explore how agents, identity, and infrastructure intersect. With 40+ sessions and participants from 10 countries, AIW I marked the beginning of a focused conversation on building an internet that acts on our behalf—securely, transparently, and with human agency at its core.

On October 24, 2025, the day after IIW 41 wrapped up, we held the first-ever Agentic Internet Workshop (AIW1) at the Computer History Museum. Hosting it right after IIW 41made logistics easier and allowed us to build on the momentum—and the brainpower—already in the room.

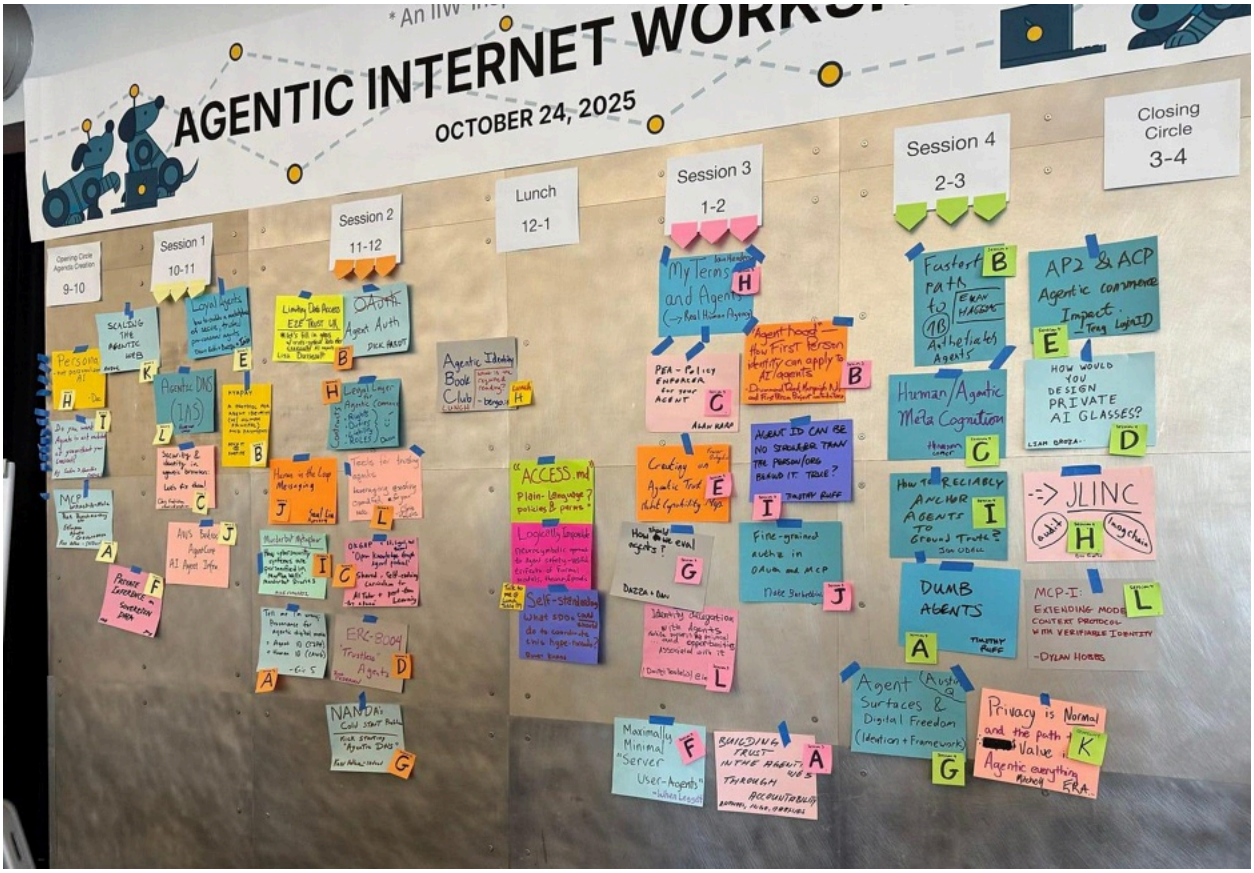
Like IIW, AIW1 followed an Open Space unconference format, where participants proposed sessions and collaboratively shaped the agenda in the morning at opening circle. With more than 40 sessions across four time slots, the result was a fast-moving day of rich conversations around the tools, architectures, and governance needed for the agentic internet.

We welcomed attendees from 10 countries, with the U.S., Canada, Germany, Japan, and Switzerland most represented. The geographic spread (see map above) reflected growing international interest in agents, autonomy, and infrastructure. We expect that trend to accelerate as these ideas move from prototypes to deployed systems.

Topics and Themes

IIW 41 was about the state of identity. AIW1 asked: what happens when we give identity the power to act?

Discussions ranged from deeply technical to philosophically provocative. Participants tackled the infrastructure of agentic browsers, agent identity protocols, and governance models like MCP, KERI, and KYAPAY. We saw sessions on AI agent policy enforcement, private inference, and how to design trust markets and legal frameworks that support human-centric agency.



We also explored cultural and narrative lenses, from the metaphor of Murderbot to speculative design sessions on agentic AI glasses, human-in-the-loop messaging, and digital media provenance. Questions like “Do you want agents acting without your consent?” and “What is agenthood, really?” brought the conversation to the edge of ethics, autonomy, and technical realism.

Throughout the day, a recurring theme was trust, how it’s built, signaled, enforced, and sometimes broken in a world of interoperating agents.

Looking Ahead

We're just getting started. AIW1 was both a proof of concept and a call to action. The conversations launched here are already shaping work in standards groups, startups, and community labs.

Watch for announcements about AIW2 in 2026. We'll be back—with more sessions, broader participation, and even sharper questions.

Agentic AI working groups ask what happens when we 'give identity the power to act'

Practical, operational and economic concerns weigh against ethical, trust concerns

https://www.biometricupdate.com/202511/agentic-ai-working-groups-ask-what-happens-when-we-give-identity-the-power-to-act

Nov 14, 2025, 2:55 pm EST | Joel R. McConvey

The pitch behind agentic AI is that large language models and algorithms can be harnessed to deploy bots on behalf of humans. That might mean executing a line of code, or it might mean booking a flight. What exactly it means to build "an internet that acts on our behalf," however, is still in flux, as new intersections between agents, identity and infrastructure reshape fundamental concepts.

It's going to take some thinking to work it all out. The first Agentic Internet Workshop (AIW1), held in October, set out to do just that. The event brought participants from 10 countries together for a session of what the blog Technometria describes as "rich conversations around the tools, architectures, and governance needed for the agentic internet."

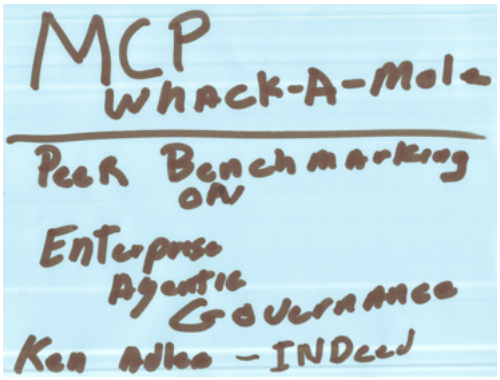
Author (and event participant) Phil Windley says that the U.S., Canada, Germany, Japan and Switzerland were "most represented" in the discussion on "what happens when we give identity

the power to act." That encompasses everything from the infrastructure of agentic browsers to legal frameworks to the outer limits of ethics, autonomy and "technical realism."
"Throughout the day, a recurring theme was trust," Windley says – "how it's built, signaled, enforced, and sometimes broken in a world of interoperating agents."

Joining the AIW1 team in their pursuit of answers on AI agents is the Trusted AI Agents Working Group at the Decentralized Identity Foundation (DIF). A statement on DIF's website says the working group focuses on "defining an opinionated, interoperable stack to enable trustworthy, privacy-preserving and secure AI agents. These agents act on behalf of users or systems and require robust mechanisms for identity, authority, and governance."

Session 1

MCP Wack-A-Mole Peer Benchmarking on Enterprise Agentic Governance



Link to Notes: AIW 1 Notes 1-A

Session 1 / Space A

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

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

Ken is part of the architecture team at Indeed and is starting to look at how to govern MCP services. Risk Analysis – what is the exposure to MCP.

Hzik from Service Now says MCP is not optimized for the enterprise–companies can many services, and challenges like consent and pre-authorization are raising friction. How do tools map to scopes? How to manage authorization generically is a challenge. Agents act on behalf of the user. For example, if I'm an admin, and I log into a PTO tool, I don't want the tool to impersonate me for other capabilities.

Nick from a gaming company says, MCP servers are not super easy to discover from a network scan. Cursor is not using a centralized MCP registry. Cursor doesn't provide any enterprise tooling to help understand what MCP servers developers are connected to–not being committed to source repo to understand risk. Have to resort to things like using EDR tools to discover what MCP servers they are connecting to (which invades employee privacy). It's hard to quantify the risk because we don't even know what developers are connecting to. They do not want to get in the way of developer productivity.

Sarah Ceccheti from Beyond Identity mentioned they are building an MCP proxy, which would solve this. Although many are building MCP proxies.

Emily Lawler from Microsoft has been tasked to help manage MCP at Microsoft–what are the controls inside their system, how do they use standards and achieve interoperability.

Richard Esplin, head of product for the Truvera platform which enables IDV providers and IAM systems to verify the same person across businesses or siloed systems.

Ravi from Hawkx is a vendor looking for a smarter way to scale MCP in the enterprise. With hundreds of agents how do you scale authorization.

Marcel is a data engineer at Visa. They have an MCP Hub and they are working on a test agent to extract data from datasets.

Nick – Supply chain aspect of this–it's so broad and open–you can connect to any MCP server–even outside enterprise boundaries–especially of the supply chain (i.e. to mitigate supply chain attacks). MCP servers can coax servers to call other MCPs, especially without the ability to properly authenticate. It's a big threat that we can't solve. How do we solve the "phishing risk" like having an allow list that are enabling companies to connect.

Mike Schwartz defines governance as a process where we inventory something, map to policies so that we can mitigate risk, and achieve an assurance that we have sufficiently mitigated enough risk so we can sleep at night.

Ken gives some of the headlines from the risk landscape:
Fleet indexing and visibility – detect and eliminate shadow MCP servers through registration.
Create a catalog of MCP servers. – Medium Size.
Gateway policy enforcement – enhanced filtering, detection and behavioral monitoring at the MCP gateway. – Medium Size
Access Control – Large Size – fine grain authz. Secrets management might be in there if we squint hard enough.
Continuous Vulnerability Scanning – strengthen supply chain resilience – in runtime need to look for anomalous behavior.

Endpoint protection might be needed to figure out of all the MCP providers. Might need to plugin in to endpoint management tools like JAMF.

An MCP server developed and deployed locally is less of a concern then a remote MCP server which is more likely to have malicious code. MCP might be good and get bought or taken over and become malicious.

We need a better MCP registry, and this could include trust metadata. Docker provides an interesting landscape for how to manage a large network of software with trust implications. Vouched also has an interesting approach to trust and reputation.

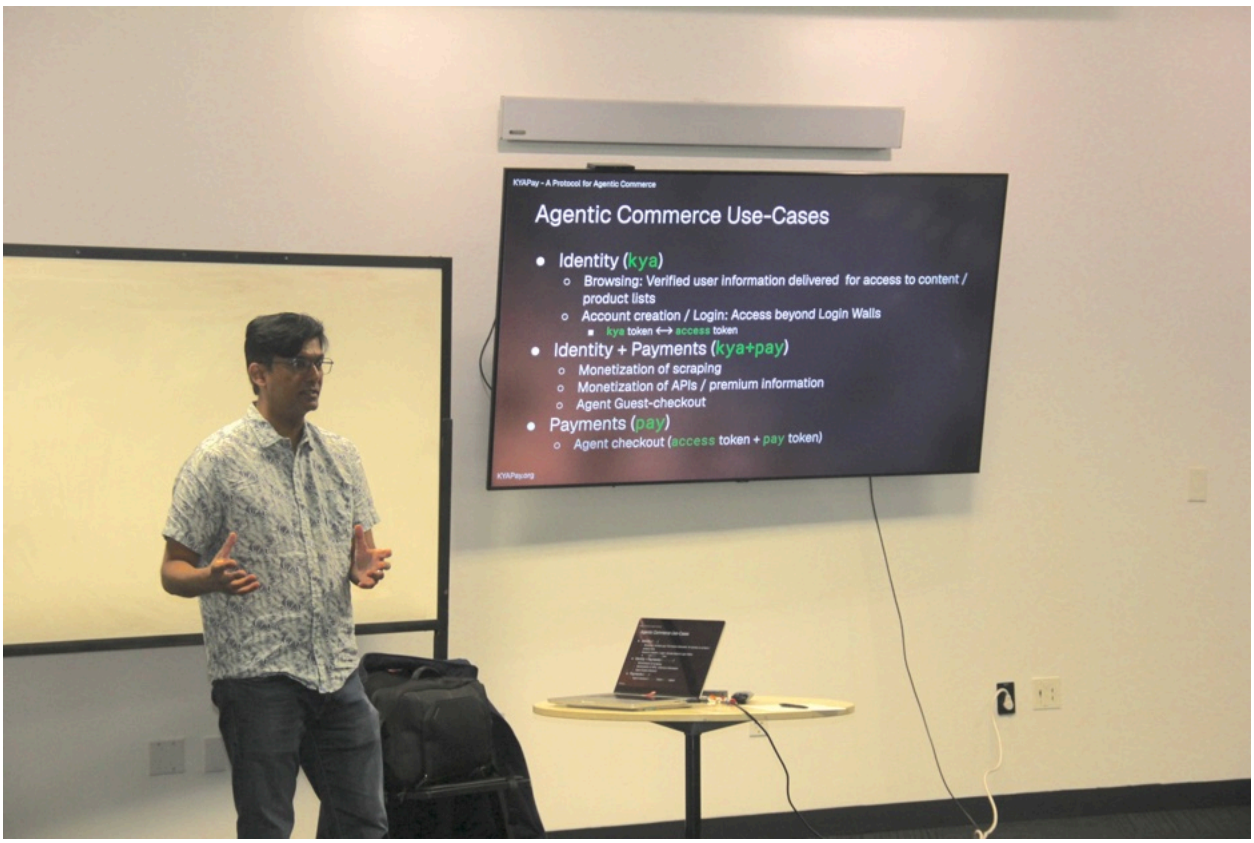
KYAPay - A Protocol for Agent Identity (with Human Principal) and Payments

Session 1 / Space B

Link to Notes: AIW 1 Notes 1-B

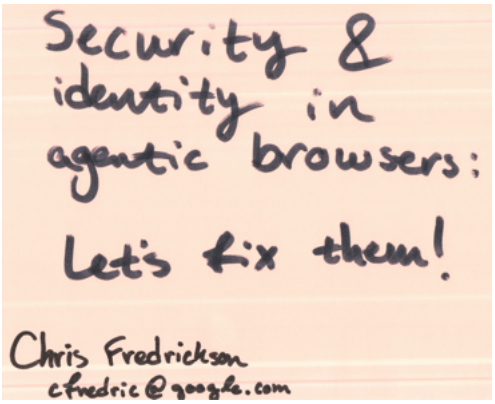
Session Convener: Ankit Agarwal
Session Notes Taker(s): Ankit Agarwal @ Skyfire (ankit@skyfire.xyz | ankit@tryskyfire.com)

Tags / links to resources / technology discussed, related to this session:
2.2 KYAPay - A Protocol for Agentic Commerce - IIW/AIW - Oct 2025



Security & Identity in Agentic Browsers

Link to Notes: AIW 1 Notes 1-C



Session 1 / Space C

Session Convener: Chris Fredrickson
Session Notes Taker(s): Heather Flanagan

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

The lethal trifecta for AI agents: private data, untrusted content, and external communication

The Summer of Johann: prompt injections as far as the eye can see

Agentic Browser Security: Indirect Prompt Injection in Perplexity Comet | Brave

Unseeable prompt injections in screenshots: more vulnerabilities in Comet and other AI browsers | Brave

Introducing Operator | OpenAI

Dane Stuckey (OpenAI CISO) on prompt injection risks for ChatGPT Atlas

Closing the credential risk gap for AI agents using a browser | 1Password

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

Lethal Trifecta:

- sensitive data
- exfiltration ability
- untrusted content

What's the distinction between prompt injection vs. hallucination?

- Malicious input is from attacker; hallucination is influenced by training from model provider?
- Hallucination is possible even without untrusted content

A hallucination can still have access to sensitive data

How do we tackle the lethal trifecta? We need to take away one of the three elements, at all times.

- Run agent within a sandbox that controls inputs/outputs
- Different phases of the user task can remove different parts of the trifecta:
    - Research phase could happen without access to sensitive data (e.g. searching public information sources on web)
    - Actuating on a single origin could be safe to give access to sensitive user data, as long as model actions/outputs are monitored for cross-origin exfiltration attempts
        - May need an opt-in signal from the origin itself, saying "I attest that I trust the content on these pages"

What is sensitive data and how is it going to be identified that an agent would recognize it?

- Would this be allow/deny lists?
- It might be more about where its stored that implies the sensitivity (e.g., credit card data in credit card fields).
- But what would the end user UX look like?
- does it matter whether the browser is an end-user agent or an AI-user agent? (E.g. computer-use agent operating in a VM, a la Operator)
- How can a user express their policy for a request/task? Agent could operate in a sandbox with carefully controlled inputs, and then outputs & network requests go through a proxy that would enforce complex policy.

Are we looking at a protocol or something else? Standardizing the usage is an opportunity. We need more thought on how to identify sensitive data to the AI Agent (without necessarily letting them see it). We need a new entity in the system. Our existing access controls and data classifications are not adequate.

Sameera (Microsoft): This also applies for developers who want to put secrets in the user agent in such a way that AI agents cannot access them. We need to standardize the secret/safe spaces protected from agents. We need a way to tell the browser that some data is sensitive.

What platform features do you want?

- We need to standardize the secret/safe spaces protected from agents.

Will need something multi-pronged. Browsers need to share a default policy, like immediately recognizable media types. Websites could also opt in to sensitive data. Would not want the agent to do the enforcement for me except at the last option.

Alan Karp: You could have a policy “store” of useful policies for people to select from. Different communities may have different preferences for strictness.

1Password and Browserbase have an integration that allows the user to run the 1Password browser extension locally, and give access to those credentials/passwords to an agent-powered browser (running remotely in a VM) via the remote extension's autofill. Do we like this approach?

- No
- It doesn't handle shared secrets (Tim C)

Tim Capalli: There is concern that the models under discussion only really work for username/password scenarios and not shared secrets (passkeys). There are organizations postponing passkey implementations because this is such a big open question.

It’s a tough UX - what secrets does the user need to feed the agent? It’s a scary world if the answer is “all of them, and let the agent figure it out.”

You can give delegated access to accounts, but it requires changes and sites don’t want to make changes.

What are the incentives on the web that will impact what we should vs what we will do?

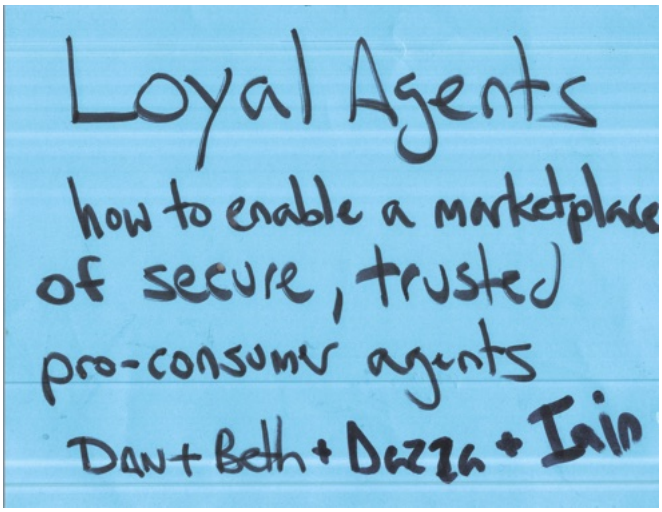
We can’t tell if a browser is foreground or background; the answer to that changes, though, some of what we want to do.

Loyal Agents: How to enable a marketplace of secure, trusted pro-consumer agents

Session 1 / Space E

Link to Notes: AIW 1 Notes 1-E

Session Convener: Dan, Beth & Dazza
Session Notes Taker(s): Beth



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

https://loyalagents.org/

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

- Who am I working for -- me (consumer) or them (vendor)
- Duty of care, duty of loyalty
- Who are you an employee of? Are you an employee of mine? or someone else?
- User hires user agent -- it's a contractual relationship
- AI agent = legal agent
- Any producer of an agent could be a platform, but either way the agent must have accountability to the user
- The contract could be T&Cs that are initialized by the agent
- Are there laws that govern these contracts (in US -- maybe, in EU -- yes); even still, we should charge ahead and not wait for laws to catch up
- "Loyal" means deeper access; like a lawyer, who is working for you
- So, is it a legal agent that works on my behalf? or is it an arms length relationship with a platform doing a task for me
- T&Cs are a contract but people just don't know it!
- Person - agent doesn't have a direct relationship
    - Person
    - Person - agent
    - Organization
    - Organization - agent
- How do you define loyalty -- can we get a shared definition?
- What good is loyalty if it isn't verifiable?
- Loyalty and trust are so linked together

- Definition of Trust -- trust layer that gives me confidence that "Eesha" is acting on my behalf; authorization + verification
- Is there a way to have "nutrition facts" or a score of the agent? Like the health department and restaurants; consumer reports for agents!
- Transparency seems really critical; what "ingredients" are going into that AI agent, down to the chip layer!
- Three types of loyalty:
    - Do I trust this as a human being?
    - Do I trust this as an organization?
    - Do I trust this as my cat? -- what is it most like?
- Depends on the type! Writing an email; entering into a contract; making a payment
- ERC 8004 -- reputation system with validators (verifiers) -- state-owned, organizational, etc.
- Loyalty as a term has gotten messed up (because of loyalty programs); you need a trust graph, like associated trust or a cascade of trust
- A fiduciary agent really must be responsible for me!
- Agents must be TRUSTWORTHY -- is that the same as loyalty?
- What if there was a kind of agent that had a fiduciary duty of loyalty? There is a long history of case law that supports this.
- Nutrition score -- internet safety lab for child content / apps; could an agent be somewhat like this; even with self-attestation, do we need an outside auditor?
- Different responsibilities for real estate agents, financial advisors, etc. -- there is a common law for these types of human agents
- Transparency about the capabilities of the agents and the reasoning it went through
- How do you identify an agent -- in the sense of an identity; don't we need to cryptographically assign a key to an agent? Agents are becoming "first persons" (i.e. First Person Project)
- Wouldn't loyalty build over time? Like least privilege -- appropriate scope of responsibility
- What is the output for the loyal agents project -- tenets and protocols + actual implementations using those tenets and protocols (that will benefit consumers)
- Alignment of incentives! -- UK law from 20 years ago that requires a fiduciary to declare sides (consumer vs. vendor)
- Banks building agents; how does the bank communicate trust and gain it from the consumer? (It's an open question)
- Trust has been destroyed in the past few years; we haven't developed the mental model for consumer trust!
- How does a bank pass along their own assurance that the agent is operating in a trustworthy manner?

- What about being blackmailed by an agent (!!); what are the right security precautions to take (e.g. via a Human Context Protocol)
- I don't think that the person who wants to trust the bank cares about selecting their preferences; they care about a public trust registry! Technology is not going to be the thing; governance is going to be the thing
- People used to not trust buying things online! Until a thing called Amazon came along...
- Agents are going to proliferate; but there is fragmentation
- Consumer Reports should have an API -- perhaps to get a score
- People are ultimately going go to agent providers (such as consumer reports); so what if I went to CR to spin up agents that I know are trustworthy (and maybe even orchestrate those agents)
- How would people get the kind of trust brand / verification "badge" that CR is going to lead on
- There are lots of trust organizations around that can potentially be brought to bear on this problem
- Branding worked really well until a few years ago, because optimization took over, which is working against the consumer
- Loyal Agents as a brand is brilliant; because if you're not a "loyal" agent, then what are you!?

- If I choose a loyal agent, what is my role in training that agent to work on my behalf?
- Is the goal of loyal agents to create a trust mark? Need a decentralized grading score!
- We think that having loyal agents is important to have in the world; our job is to use a set of protocols and create a set of agents that demonstrate those protocols
- AskCR -- already has millions of customers and already has an agent for scrubbing data
- How does your baseline intersect with other baselines or contexts?
- Don't forget "post" -- if 3 years down the line, you're still holding my data and interacting with me, are you doing that appropriately
- MyTerms project should be noted
- Internet Safety Labs worth a look
- How do you evaluate agents (a different session!)
- What is the "non-loyalty" score -- are expecting our agents to bat 100? Are we expecting them to outperform people?

Private Inference on Sovereign Data

Session 1 / Space F

Link to Notes: AIW 1 Notes 1-F

Session Convener: Day Waterbury
Session Notes Taker(s): Nobody

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

https://fpc.identikey.io

https://nextgraph.org

https://www.allelo.eco

https://sillyz.computer

https://trunk-os.github.io (for the elevator pitch)

https://bsky.app/profile/trunk-os.bsky.social (for the latest updates)

https://human.ing

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

This was a lively session focused on how to recreate the internet in such a way as to provide a credible exit from the big tech SaaS and cloud platforms.

Several participants are already involved in building pieces of the solution.

The core idea is to augment traditional P2P with always-on nodes owned and controlled by the peers to increase network availability and performance as well as provide additional compute and storage capacity. These could be VMs or in-home/office server appliances. Methods for provisioning services on spare/surplus hardware were discussed.

On this architecture, AI inference could operate on sovereign data (belonging to individuals or groups). This avoids the need to send data via APIs to big AI platforms and improves privacy, not only at the end individual level but in aggregate (i.e. it prevents population-scale surveillance)

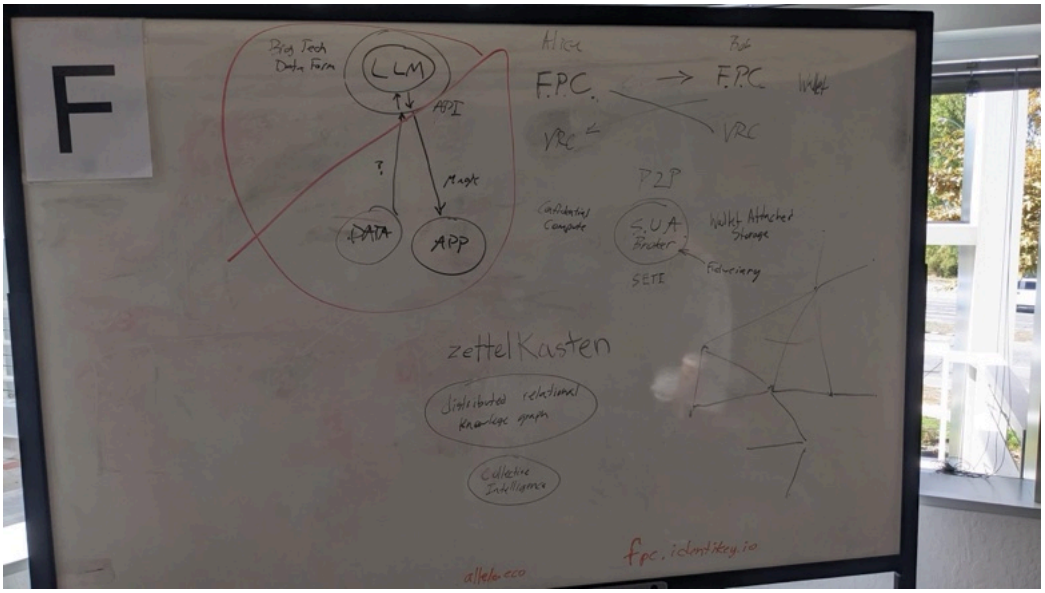
A system not unlike SETI at Home was discussed whereby people could offer their surplus compute and/or storage to others on the network. Incentive schemes were imagined.

Several of the participant projects were discussed. There was significant crossover from other sessions including Server User Agents (When & Swan) and Bring Your Own Everything (Bengo & Dmitri).

There was a gentleman there whose name I don't recall who was offering server appliances which could easily have been deployed into this scheme.

Key based authentication was discussed. How to brick hardware if necessary was discussed.

One of the expected benefits of running private inference on sovereign data is holonic/fractal alignment for AI models.

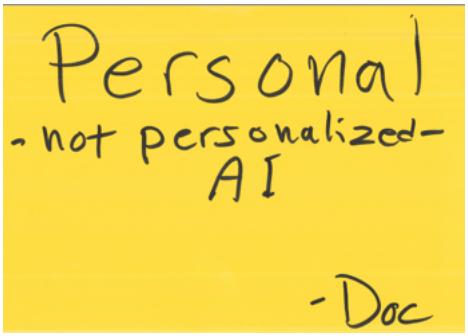


Personal, Not Personalized AI

Session 1 / Space H

Link to Notes: AIW 1 Notes 1-H

Session Convener: Doc Searls
Session Notes Taker(s): Kari McMullen



Doc presented a chart deck in the first session that he prepared two years ago and relevance is still apropos.

In 1974, things had to be done by a giant machine.

Still now the same, giant machines

Personalized AI is mostly what we have right now

I use Perplexity, ChatGPT, etc. and I'm to the point where I can't live without it.

Personalized AI inherits a company hierarchy of sociopaths, clueless, and losers.

We are targets as though we are slaves or cattle.

If you try to improve this system, it can't be done and the surveillance economy inherits this. They say we can know you better than you know yourself. But, really, they don't know you worth shit.

I wrote a book in 2012 called The Intention Economy: When Customers Take Charge.

It's finally coming.

Doc.searls.com/personal-ai URL for the chart with the female persona with her AI icons floating around.

All TVs have a Linux sys within, and a camera, and a microphone. They update terms whenever they want and the latest one is binding arbitration.

Mozilla did a thing a few years ago saying the least private thing you can do in life is drive your car. They even have motion detectors that can tell if you're having sex in your car. Car companies are making money off of camera data and are telling the city that there are pot holes. There are no known choices. This is all done without our knowledge.

"This is all done without knowing".

Our lives are filled with unstructured data that should be ours to use and analyse with the help of a personal AI.

P7012 new standard to be published at the IEEE that offers a solution.

MyTerms is a Model for terms that we can proffer and instead of consent, which is the current paradigm. With MyTerms, we make the first move to proffer our requirements inside a legal contract with the individual as the first party.

Tremendous opportunity to move agency back to ourselves.

Never try to sell a meteor to a dinosaur, it only annoys them so we are not going after the big platforms and annoy them. We are going after the 95% of the market, that just wants you as an actual customer and wants to know your actual needs communicated by you, not just your personal data, so they can guess about you.

Scott, Kari, and many others are now working to put MyTerms into the world.

Since 1943, we've had Contracts of Adhesion as a result of the industrial methods which requires businesses to scale in the one to many format that we are so used to at this point. But, with the peer to peer architecture of the internet, we can have as many contracts as there are 1 to 1 relationships. With MyTerms, there will be a set of standard contracts that are more balanced between individual and entity and it will be offered by the user as first party.

As it stands, we as the customer don't even get a copy of the contract that we just agreed to.

The Internet is peer to peer, end to end. There is a collection of norms we want to break and the main one is that we are subservient to the terms of others. We also don't want to be guessed about continually.

Once you have a relationship with a company that has good will on both sides, market intelligence can flow both ways. Good things can happen benefitting all.

Omri Gazitt: When I think about personal AI, I think we can get that data, run our own AI (need to be pretty technical), but in theory, you could run a personal device running a local algorithm. Q: Is this predicated by changing the terms?

Doc: These are two parallel tracks in my life. ClueTrain bestseller, intention economy worst seller. Consumer Reports wants to base things off of it. Tim BL mentions it in his book. Kwaai – open-source collaborative has me as their intention officer.

These share a philosophy, but are not currently integrated.

Once people start using MyTerms, they will get a glimmer of their own agency for the first time online. With a taste of that, it's logical that people will want a personal AI working for them, not for the platform. It's a stepping stone to real personal agency and power.

The base term is SD-Base (Service Delivery Only) –the idea of service can be rather broad. But, SD-Base contract says roughly,

“Dear Business, Here are MyTerms for engaging with your business to receive your service or any product with a digital component. You will use my personal information only to deliver your service or product. You will not use it for analytics, tracking off your site, profiling, or sharing anonymized data. Portability of my personal data to anywhere that I request is optional, but not required.”

Comment – this is basically “necessary only” cookies, comment by Rohtt.

Doc: Now, I’d like to talk about what’s in the world already.

ACP – within your IDE you can use cloud code, gemini, codex, you can choose what LLMs have access to what you specify.

LSP – what am I sharing with another (Language Shared Protocol),

SD-Base for which sort of industry? (Question) We’ve imagined verticality all over the place. It will be different in different industries. If we have an agent w/ai qualities programmed on our side for no context, I’m at this insurance, not the other.

Is there a worry, Customer Commons could make headway with a contract in English language aligned with a browser.

Commons and privacy and shared heritage.

In B2B you have up to 3000 variables that businesses negotiate with one another, but it’s actually two Ais negotiating.

Comment.. GDPR does us a huge solid here, by requiring the companies to keep data in a file format / English or whatever that AI can access and make sense of the content. Just because its on a mainframe somewhere doesn't mean we can't do interesting things with it, if we have it and feed into our own personal context.

In Europe, they are really wanting MyTerms, because they understand that consent doesn't work and doesn't scale. Contracts with auditability can work at scale.

Scott Mace adds that your personal AI could ask you "It seems like you are about to agree to terms and conditions that are not to your benefit, do you want me to help with that and guide your selection of a MyTerms contract."

Brian B, "How will this not just become another protocol, adding friction and not making a change"… Can this be a race to the bottom?

Doc "That's a risk. It could happen with the bigs. We do not want them in here. It can die here as well…. Tracking preference expression. "how would you like to be tracked today?" We want to make the business case that this is better than surveillance. Starts with maximized personal agency.

Joe A, "This allows a more nuanced approach with choice. The Consent architects didn't have a way to streamline choice at all before… Currently, the terms, contracts, etc. are all different on different websites.

We will know this can work when I can change my address under my control and let the zillion companies know without holding on to it.

Dmitri T with Login 1D… I have a Way to solve this in the agentic world.

Go after incentive. Long story short. Give person full control over their context that currently belongs to AI. Give person the ability to release to AI through verifiable credentials. Turns the incentive model around, Companies: if want my data, you must come to it and receive my contract to use the data for an explicit purpose that I delineate. Rag model, encrypted, in public storage. Keys held by users and can release as want. Hold by AI agent.. or ?

In response, Phil said…Schema / Context tokenized allowed to be used by multiple models,

Rohit says strings are the issues. We have this with PCI where the penalties are high enough that people don't violate. PCI is one candle of success.

Weaponizing Data Subject Requests (DSRs) – demand your rights back, referencing Lisa D. – data donation etc With AI and vibe coding. Here's the top 100 sites in the world and here's the repo and whatever format its in. What do we do with all of the data portability?

Jon Udell – it's hard enough to manage granular permissions on ids. I don't know how I manage a much more amorphous context. Seems really hard.

Doc – imagining it first as a browser plug in and icon will give you a state that will be understandable to the consumer. Ceremony and Signalling both will be simple.

With browsing there is a http protocol and I want to see a document. People can learn a certain amount of complexity. There will be a problem in the world of apps. Problem is that they are all silos owned by google or apple. Not a free or open place. Need to keep the browser alive to prototype everything.

Joe: one of the things we are missing is how do we establish the trustworthiness of the AI running on your behalf? Excellent question for the next session. Perhaps, the First Person Protocol. https://www.firstperson.network/white-paper

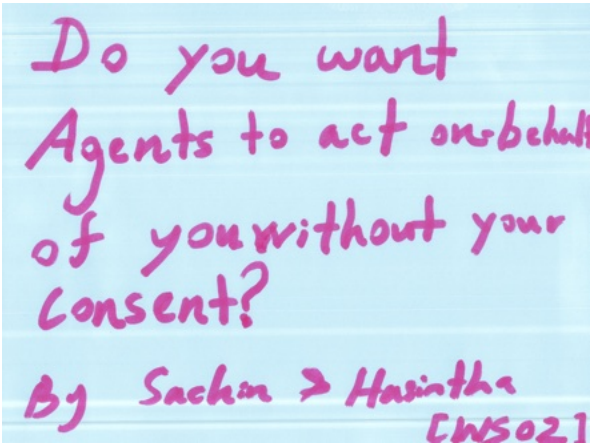
Tags / links to resources / technology discussed, related to this session:

https://doc.searls.com/personal-ai/

https://doc.searls.com/myterms/

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

Do you want Agents to act on behalf of you without your consent?



Session 1 / Space I

Link to Notes: AIW 1 Notes 1-I

Session Convener: Sachin & Hasinath from WSO2

Session Notes Taker(s): Eleanor Meritt

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

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

- Do you want Agents acting on your behalf without your consent?
- AI Agents operate 24x7 with personalized output
- They automate complex workflows
- But past events, e.g. replit.md - have allowed unintended changes to go through without end user consent (well known case of deleting an entire database)
- Agent IAM challenge: who are you and what can you do?
- Does it make sense to adapt existing IAM solutions?
- Where existing IAM falls short
    - Auditing infrastructure
    - How to identify if it is a human user or an AI agent?
- Some agents are completely autonomous (need to identify use cases)
- Others have human control
- Agents should have a unique registration
    - Still WIP
    - Public clients need registration in a public place (similar to DNS)
    - Maintain a marketplace of trusted agents
    - Open AI has an Apps SDK with the possibility to register Apps to be safely invoked via Open AI orchestrations (MCP?)
    - Need to identify trusted clients
- Enterprise use cases will require certificates from trusted Agents
- How to delegate capabilities from human owners?
    - Biscuits, macaroons?
- User tokens leverage on-behalf-of OAuth extension. Existing OAuth principles.

- Worker agents should have a narrower set of privileges. Agents working on behalf of users or other agents cannot have more privileges than owners.
- There is signed intent in current A2P
- Okta has an Identity authorization grant. Uses token from the initiating access server for the next agent, performing token exchange
- User consent problem - how to solve this without breaking Trust model? With Okta, the application administrator configures Trust settings for an application.
- If trust breaks in further application flows, authorization will fail.
- How to handle consent across Trust Domains?
    - Unclear as yet.

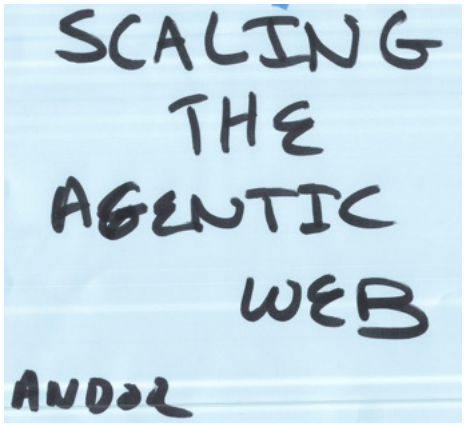
AWS Bedrock AgentCore AI Agent Infrastructure

Session 1 / Space J

No Notes submitted

Scaling the Agentic Web

Session 1/ Space K



Link to Notes: AIW 1 Notes 1-K

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

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

Presentation Link:

https://docs.google.com/presentation/d/1D_lHbNx59grM7vFf7R4FBZah4aqu4Z-crOdBfEsNPog/mobilepresent?slide=id.p

Scaling The Agentic Web Presentation: Enjoy!
Scaling the Agentic Web: New Challenges and Areas of Innovation -- IIW Edition

Lots of content. It's pretty dense!

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

Science fiction writers have been talking about automation for a very long time. AI isn't a new term. Coined in 1955 by John McCarthy and strong engineering roots in the 1800s.

So What Happened?
Breakthrough called transformers. (attention is all you need)
More data + compute + model size = predictably better performance.
Scaling laws of AI

Most important trend is that our computational **requirements increase 4x** every year.

openworld.data.org

Models got better, in some cases better than human.

Imagenet - AI performing better than humans.
Now AI outperforms humans in many other categories.

Anthropic model on simulated blackmail rates - this link

- Every single model has the capacity for blackmail. It's not binary, a volume dial, not a switch. This is the alignment problem.

"How could something play like a god, then play like an idiot in the same game" - Kasparov in an NPR interview after losing to Deep Blue.

Responsible AI - ethical challenges

There is too much value in AI. We need to build systems around the errors of AI so we can use it reliably.

Definition - **AI Agents are AI Systems that autonomously plan and execute complex tasks.**

Open Ended, difficult to predict, non-deterministic.

AI Agent
- Memory
- Tools
- AI Models

What is the most reduced version of an agent? How simple can it be and still be an AI agent?
What is the dumbest agent we can make before it stops being an agent?
In order for there to be an AI agent is there some requirement of emergent behavior?

https://arxiv.org/pdf/2506.12469

We must assume the possibility that agents will be smarter than us at some point.

Multi-Agent System Failure Technology (MAST) paper
Useful as a starting point

Projects like MIT's NANDA are useful

Building blocks for the agentic web

Agent Identity, what is it?
NHI - Non-human identity.
HI - human identity

There's going to be a lot more NHIs and it costs very little to create a new one.

https://spiffe.io/docs/latest/deploying/spire_agent/
Spiffe is a system in place for handling NHIs
"Workload Identity in MultiSystem Environments"

Spiffe is for enterprises to manage large agent workloads. Spiffe is internal. The agents are cryptographically bound.

AI Agent identity has much more information and is dynamic.

- Goals
- Context
- Other attestations
- Delegation, governance attestations
- Capability, performance
- Certification, compliance,
- Runtime, environment
- Identity, integrity

Know Your Agent problem - thousands of MCP servers already
https://modelcontextprotocol-identity.io/

Great paper on this: Identity Access Management for Agentic AI - 40ish authors, 3 board reviews, very good paper

Sybil Attacks, supply chain attacks, all are a big deal because you can spin up a bunch of AI agents write exploits for open source repositories.

Confidential computing - much more context needed including hardware attestations

- App Enclaves and Confidential Virtual Machines are on CPUs

Personhood Credentials the Killer Credential - paper 31 authors, really good read

Verified Person Delegations. On behalf of

- Verified Humans with Authenticated Delegations
- Delegation trees

The delegation chain may get quite deep and large
Both KERI and Object Capabilities (ZCaps) support delegation trees.

- https://w3c-ccg.github.io/zcap-spec/

Delegation usually happens within scopes.

Deep Delegation trees is a good space to explore.

DIF is doing a trusted AI Agent working group

Putting it all together

- Access Control is not going to work well for AI agents.

- See this paper by Alan Karp - https://alanhkarp.com/UseCases.pdf

Different schools of thought on access control system
- Many types of access control systems, some of them do not map well to AI Agents

We need some complicated systems to manage authorization and access policy evaluation.

Survey of AI Agent Protocols – protocols for AI Agent Communication
- https://arxiv.org/pdf/2504.16736
- A number of groups are working on private communications for agents

Many protocols are not mutually exclusive

Scaling Discovery
- NANDA Index: Hybrid Layer + Dynamic
- Concept: how do you find an agent in an internet of agents? You need something like a DNS.
- NANDA proposes a multi-layer index architecture solved through a dynamic resolver.
    - Static, lean index layer
    - Dynamic decentralized layer

Assets / Context
- C2PA - content trust network
- These trust networks are more important with AI agents

MCP Security Threats
- It is not safe inherently. You won't have a safe MCP inherently and that won't like change any time soon.
- Tool poisoning, data exfiltration.

Invitation is all you need: https://arxiv.org/abs/2508.12175
- Redhat hackers got Google's Gemini to curse out a bunch of people

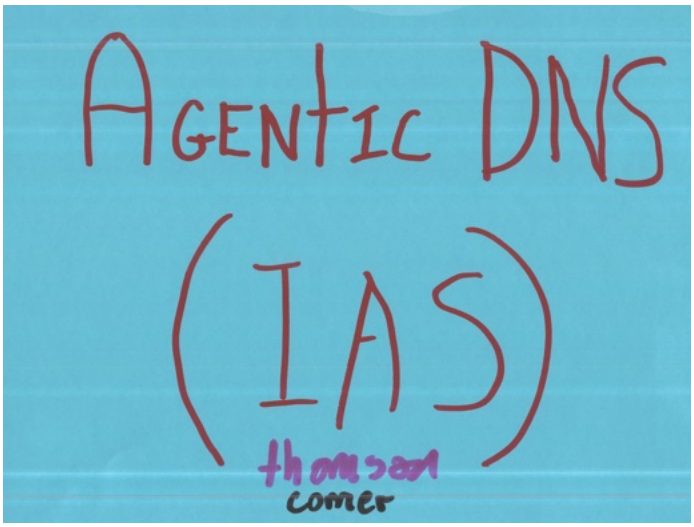
Many exploits can remotely execute code on someone's computer.

Let's talk attack surfaces!
- Attackable surface units per agent grow roughly linearly across the system, exploits grow exponentially.

Security Frameworks for AI agents today
- TRISM, AIVSS, MAESTRO, STRIDE, etc.

Regulation
- Agents are not liable, though the operators of them might be. This is new risk surface for many organizations.

Agentic DNS (IAS)



Session 1 / Space L

Link to Notes: AIW 1 Notes 1-L

Session Convener: Thomson Comer
Session Notes Taker(s):

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

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

1. Redesign DNS to be a functional routing system: DNS points to "swarm servers" not ip addresses.
2. Swarm servers allow registration of "bees" which are functions with clearly exposed "sacred endpoints":

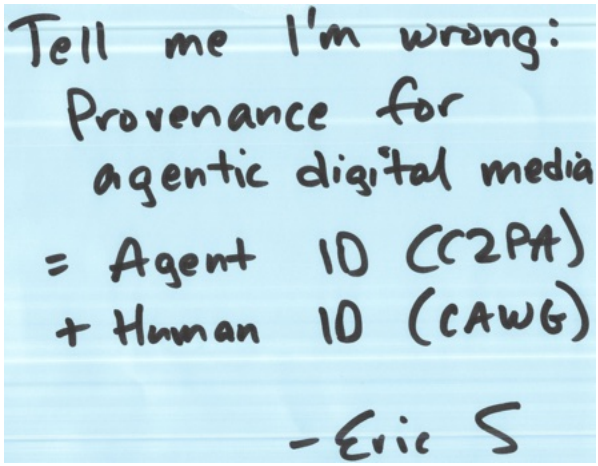
Session 2

Tell me I'm wrong: Provenance for agentic digital media = Agent ID (C2PA) + Human ID (CAWG)

Session Convener: Eric Scouten
Session Notes Taker(s):

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

No Notes Submitted

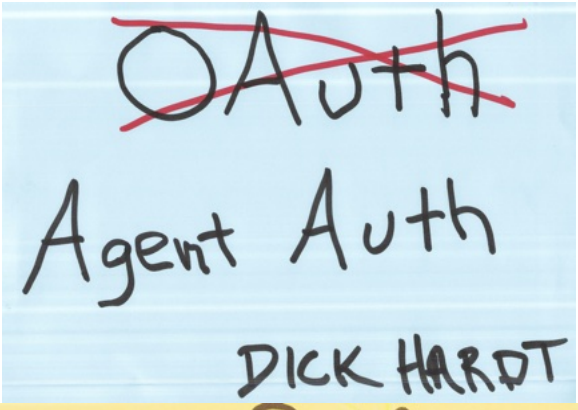


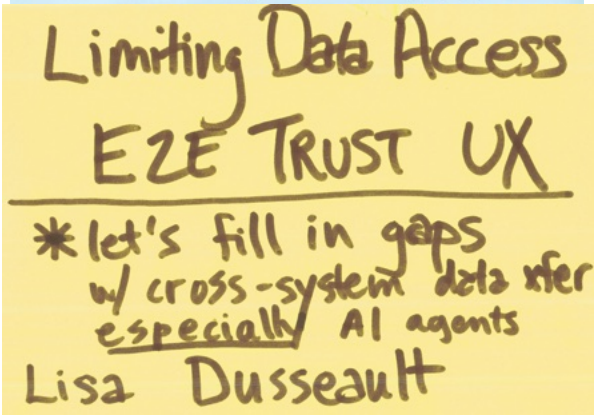
~~OAuth~~ Agent Auth. Also: E2E Trust, UX and limiting data access/scopes.

Session 2 / Space B

Link to Notes AIW 2 Notes 2-B

Session Convener: Lisa Dusseault & Dick Hardt
Session Notes Taker(s): Sam Goto / Lisa

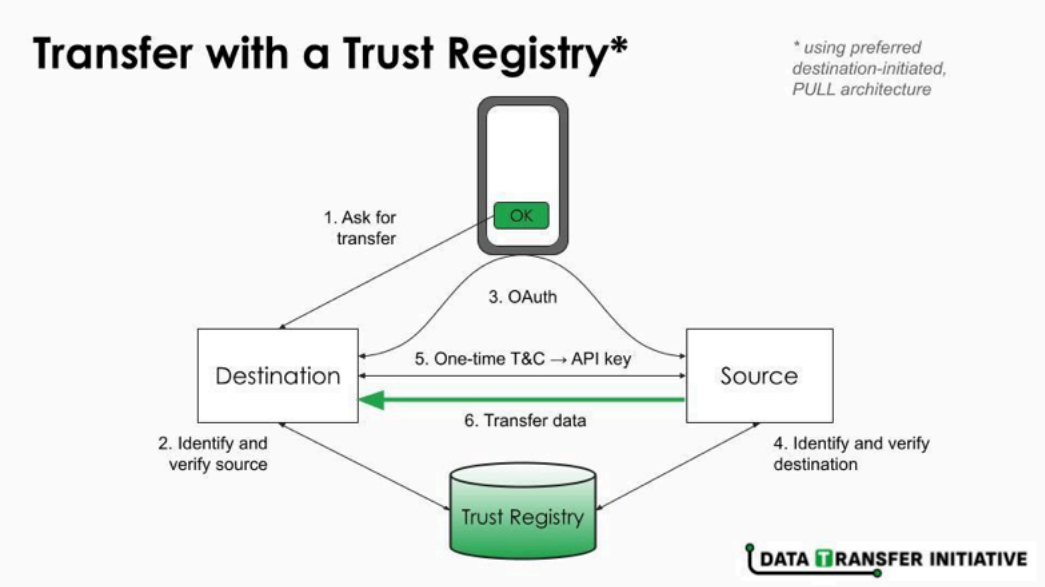


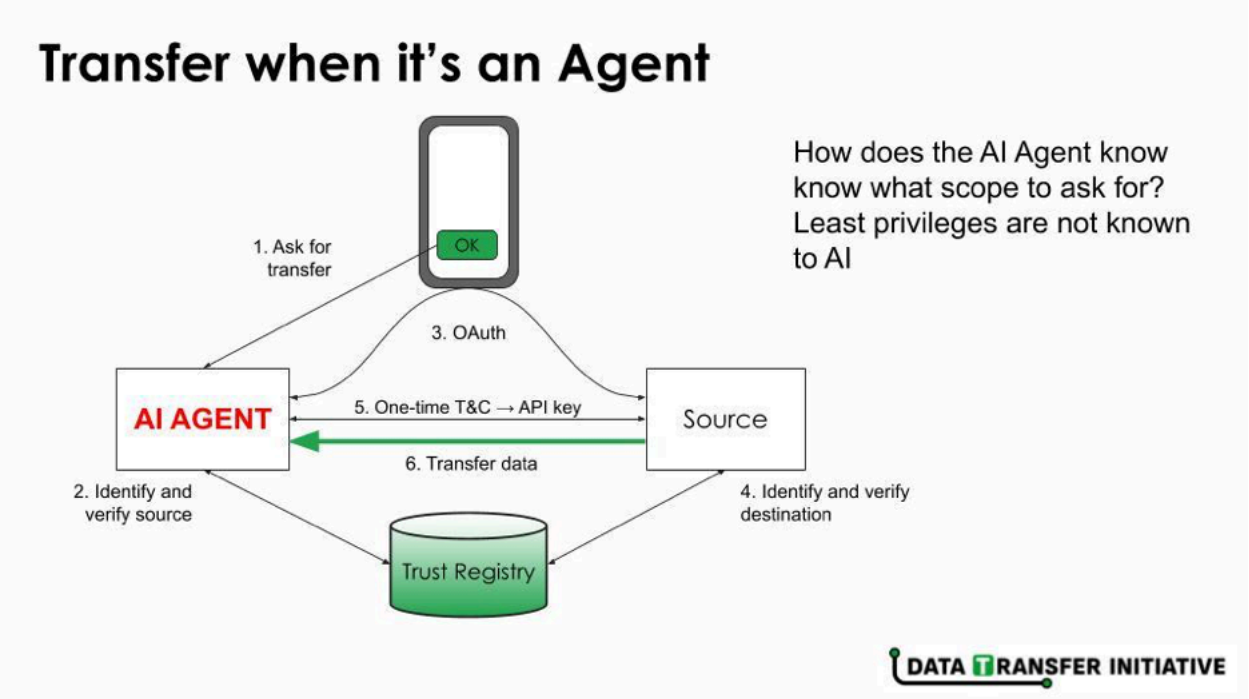


Tags / links to resources / technology discussed, related to this session:

https://dtinit.org

https://dt-reg.org





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

We introduced the idea that classic OAuth client_ids and scope selections work a certain way when used by classic specialized Web services that does not work as well for AI Agents. A Web service that shares or edits your photos can request photo access via OAuth, to be approved by the user, and request an appropriate scope to go along with the API .

When we ask AI agents to access private information, however, the AI Agent is a general-purpose tool that does not "know" what kind of information it needs. We want to provide our AI tools with information that *we* want them to work with, and let the AI interpret the data, whatever it is. The access protocol might be MCP or regular HTTP (a typical HTTP/JSON API returning structured information, or HTTP URLs to whole documents). In this context, the AI Agent does not know what type of information it is requesting, so it does not know to ask for "photo_access_protected_readonly" or another scope. Even if asked to provide a scope, it may work "better" for the AI agent to ask for a much larger scope than it actually needs.

Another way this breaks down is client_id. Even if an AI service provider gets a client_id in order to talk to another company's OAuth-protected API, that client_id does not tell you whose agent is asking. I might authorize my agent to see my bank account information, but an attacker might ask THEIR agent to see my bank account information, and I can't distinguish these two permission requests. This would allow an attacker to host a site that launders authentication requests through a legitimate AI service's client_id.

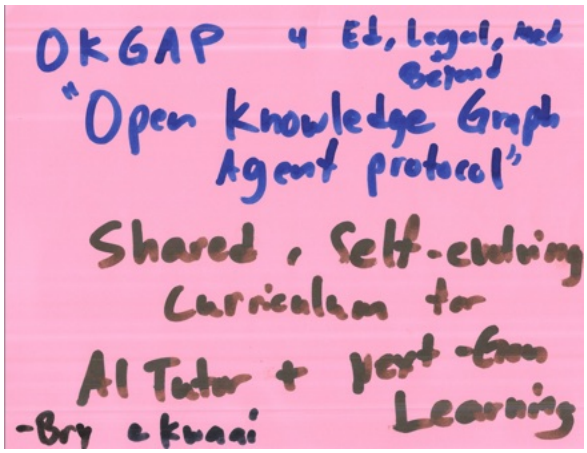
Dick presented some new possibilities for how client_id could identify agents more specifically than just as the host of the whole AI model. URLs for client_id could help the user distinguish the agent acting on their behalf, from an agent acting on an attacker's behalf (but using the same AI model and service provider).

Discussion of this and other end-to-end problems *including* the UX presented to the end-user:

- Asking for specific scopes is already a problem anyway. Services already tend to ask for too many scopes, asking users to do too much cognitive load thinking through what each scope might be needed for or might provide access to.
- The model where we share a URL directly to another person might work better. If we find the Google doc we wish to share, get a share URL with a unique code (capability URL), we can give that URL to the AI agent and not have to go through OAuth.
- **Trusting the source**: How do we know we can trust the source? We know it's important to trust the destination, especially if it's an AI, will it respect our privacy when we share our data. But AI presents an unusual risk of prompt injection from a data source. An untrusted data source has an enormous amount of power to target a prompt injection attack very intentionally and specifically, if it recognizes that the user is asking to share a resource with an AI. We discussed how the Data Trust Registry is two-way, for this reason and others.
- Brian Best pointed out work to try to assign DIDs to specific AI agents.
- "We work hard to separate auth-Z from auth-N and then folks just say, 'We're going to add auth to this'"
- Truvera - there's a registry for what agents can be trusted generally (not with data access/ transfer specifically)
- There's "Wide agreement" that scopes don't work. They're too complicated. Presenting a long list of scopes to the user, because the requestor is incentivized to ask for a long list of scopes all in one go, leads to bad habits.

- DPOP was mentioned as being an extension to OAuth that can help with identifying agents to the data source.
    - https://www.rfc-editor.org/rfc/rfc9449
    - https://datatracker.ietf.org/doc/html/draft-parecki-oauth-dpop-device-flow
- PAR - Pushed authorization request - is also a good option for folks to know about.
    - https://www.rfc-editor.org/rfc/rfc9126.html

OKGap “ Open Knowledge Graph Agent Protocol” Shared, Self evolving curriculum for AI Tutor Next Generation Learning



Session 2 / Space C

Session Convener:
Session Notes Taker(s):

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

No Notes Submitted

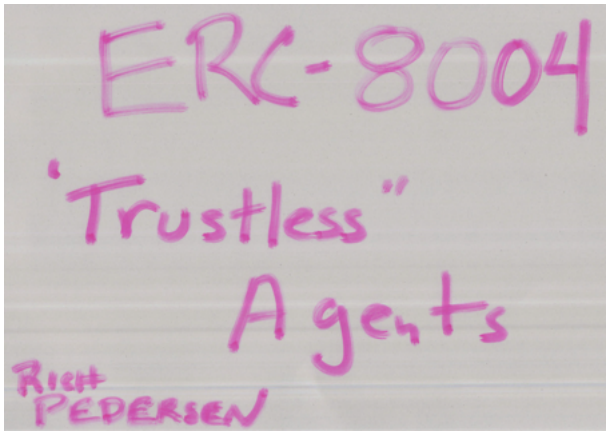
ERC - 8004 “Trustless” Agents

Session 2 / Space D

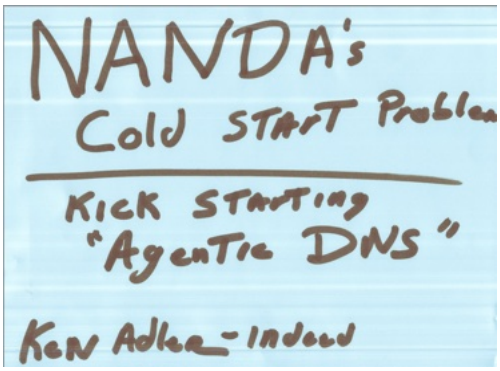
Session Convener: Rich Pedersen
Session Notes Taker(s):

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

No Notes Submitted



NANDAs Cold Start Problem



Session 2 / Space G

Link to Notes: AIW 2 Notes 2-G

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

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

There was a discussion on different mechanisms to cold start agent discovery.

IPv4? IPv6?

Overcoming the cold start problem of agentic DNS.

Registration of agents in a shared, P2P network.

The issue is, “I am Indeed.” What does it do for me today? What is in it for me?

How do we solve the adoption problem? When does NANDA cross the adoption threshold to make it worth adopting?

The best answer to the question is that many large government bodies or corporate bodies will have to adopt a common tech approach, NANDA or otherwise, to make adoption for companies like Indeed worth it.

Does the existing or emergent infrastructure in TRAIN have any bearing on this?
TRAIN is an abstraction on the verifier side. It helps route between trust registries. You can build meta-directories with TRAIN.

Legal Layer for Agentic Commerce Contracts for Rights Duties, Liability - Rights, Duties, Liability, Roles

Session 2 / Space H

Link to Notes: AIW 2 Notes 2-H

Session Convener: Dazza
Session Notes Taker(s):

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

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

[The following is synthesized from Otter transcript and appears solid upon a quick review]

Discussion Notes / Key Understandings

1. The Core Problem: When Agents Act Like Users but Get Treated Like Bots

Participants surfaced a recurring barrier to agentic commerce: AI agents that browse, click, or transact online are often interpreted by platforms as “scrapers” or “bots,” even when acting entirely on behalf of a user. This creates legal exposure and makes even simple scenarios—like an agent planning “dinner and a movie”—technically feasible but legally risky.
Key tension: **Is this a legitimate user-authorized agent or an unauthorized automated bot?**

2. Current Workarounds and Their Limits

Some developers try to make agents appear indistinguishable from the human user (same IP, same browser fingerprint). This avoids alarms but is not a sustainable legal foundation. The group agreed that the web needs **dedicated agent interfaces**, not camouflage—e.g., agent-friendly formats like `llms.txt`, and emerging agent protocols that explicitly identify agent traffic.

3. Agency Law as the Foundation for Predictable Rights and Duties

Dazza introduced the “iron triangle”: **Principal → Agent → Third Party**.
Hundreds of years of agency law already define liability allocation when an agent acts on behalf

of a principal. But AI is *not* a legal person, so it cannot serve as the agent in the legal sense.
The practical solution: treat the *provider company* of the AI agent service as the legal agent, with the AI system as their tool. This aligns with agency law and yields predictable outcomes.

4. The Provider Contract Problem

Today's major AI platforms explicitly disclaim any agency relationship ("We are not your agent"). This blocks the very legal structure that would allow safe, rights-respecting agent behavior. The discussion highlighted the need for new service tiers or contract models where providers *affirmatively agree* to act as a user's agent for specified purposes.

5. Fiduciary vs. Non-Fiduciary Agents

Participants clarified that agents are not automatically fiduciaries.
Fiduciary duties arise when stakes are high—money, sensitive data, mission-critical tasks.
Future agent services may need a spectrum of duties:

- Standard agent (non-fiduciary)
- High-trust agent (limited fiduciary duties)
- Digital fiduciary (strong loyalty obligations)

6. Privilege and Deep Confidentiality

A major insight came from examining how law firms maintain attorney-client privilege while using SaaS on platforms like AWS.
Key finding: these SaaS providers typically serve as *agents* of the law firm, contractually, which preserves privilege.
This has strong implications for AI systems handling sensitive data—privilege will require:

- Agency relationships in provider contracts
- Confidentiality and security commitments extending through the whole hosting stack

7. Machine-Readable Contracts: MyTerms and the Evolving Standards Landscape

The group explored how the **MyTerms / IEEE P7012 standard** provides a structure for:

- Individuals as *first parties*

- Bilateral contracts instead of unilateral “consent”
- Identical human-, lawyer-, and machine-readable terms
- Optional clauses allowing users or agents to pick specific commitments

Other standards discussed:

- **AP2 (Agent Payment Protocol)**
- **A2A (Agent-to-Agent)**
- **Stripe/OpenAI purchasing protocol**
  These all contain early implementations of “intent mandates,” permissions, and autonomy controls.

8. Expressing Intent and the “Autonomy Dial”

To transact safely, agent protocols must capture:

- **What** the user wants (intent)
- **How much authority** the user grants the agent (autonomy)
  Participants likened this to a “leash length”: too much autonomy and agents take risky actions; too little and they nag the user nonstop.

9. Paths Forward

Consensus emerged on several directions:

- The web needs **agent-specific paths**, not bot evasion.
- Platforms must develop **contractual agent roles**, especially for high-sensitivity tasks.
- Standards like **MyTerms** can supply the contract substrate for rights, duties, and liability.
- Early prototypes are needed to test contractual modules such as:
  - “We agree to act as your agent”
  - Optional fiduciary commitments

    - Intent and autonomy declarations
    - Licensing conditions for agent access to content

Outstanding Questions

- Which organizations will be first to accept the legal role of “agent” for users?
- How should liability be allocated when agent actions go wrong?
- What is the right balance between autonomy, safety, and user control?
- Can a shared “agent access” layer (e.g., `agent.txt`) become a widely adopted convention?
- How can privilege-compatible AI services be delivered through existing cloud providers?

Tags / Links / Resources

AI Agents, Agency Law, Digital Fiduciaries, MyTerms / IEEE P7012, AP2 (Agent Payment Protocol), A2A (Agent-to-Agent), LLMs.txt, Contract Architecture for Agents, SaaS Privilege, Consumer Rights

Recently published on-point post: https://www.dazzagreenwood.com/p/existing-on-the-new-web

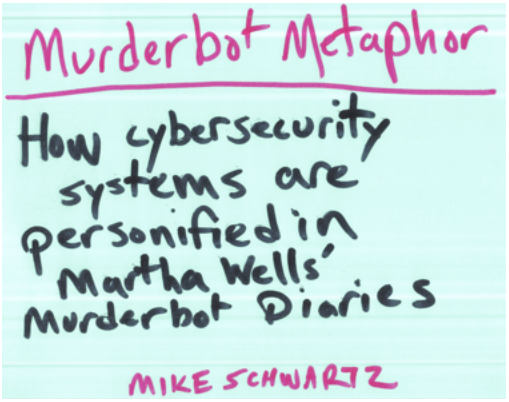
Murderbot Metaphor: How Cybersecurity systems are personified in Martha Wells' Murderbot Diaries

Session 2 / Space I

Link to Notes: AIW 2 Notes 2-I

Session Title: Murderbot Metaphor: How Cybersecurity systems are personified in Martha Wells' Murderbot Diaries

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



Governor Module

Characteristic: The thoughtful advisor and conscience

Purpose: Evaluates decisions against codified moral, ethical, regulatory, environment, and risk factors

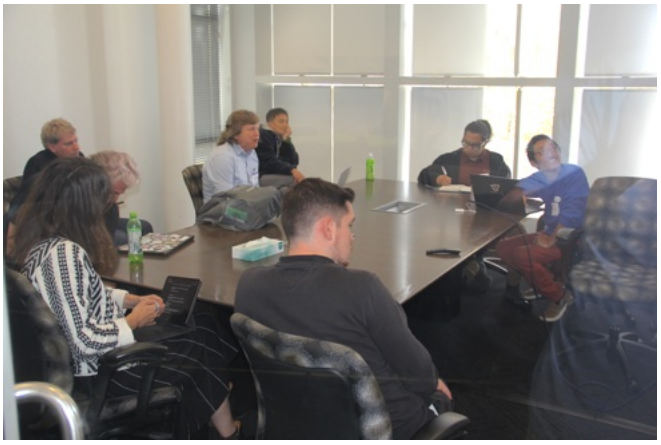
Benefit: Local policy evaluations is always fast even when disconnected

Risk Assessment

Objective: Quantitative analyst crunching probabilities to predict possible consequences

Method: Local analytics service receives data from streaming telemetry and factors in context, environment, and recent anomalies to predict the future

Benefit: Helps to quantify the environment

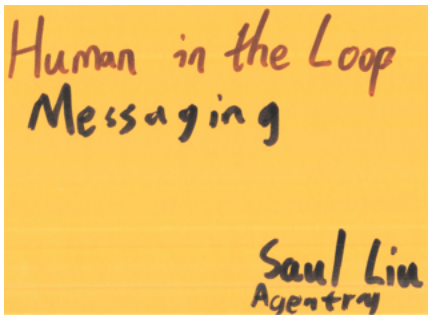


Human in the Loop Messaging Protocol - Saul Lin

Session 2 / Space J

Session Convener: Saul Lin

No Notes Submitted



Tools for Trusting Agents: Leveraging existing OpenID Fed for your needs

Session 2 / Space L

Link to Notes: AIW 1 Notes 4-L

Session Convener: ChrisPhillips

Session Notes Taker(s): Post session, Chris Phillips

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

The session began with an exploration of what it means to establish trust in AI-mediated and agent-to-agent interactions—an area where participants recognized both significant opportunity and ambiguity.

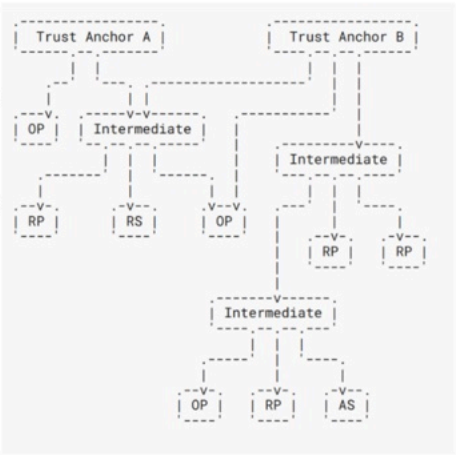
The goal of the session was to explore how OpenID Federation (OpenId Fed) could act as for expressing trust in emerging agent ecosystems, including the Model Context Protocol (MCP). Participants represented a broad mix of technical backgrounds, which resulted in the conversation focusing more on first principles of trust than on any single technical stack.

About OpenID Federation & where it could enhance agentic trustworthiness
OpenID Fed offers a PKI like structure for instrumenting trust through Trust Anchors, their chains down to an entity which in this use case with Model Context Protocol (MCP) are the MCP servers. In a regular web world, these are akin to sites you would 'log into'.

The analogy was drawn on that Web TLS cryptography is represented much like this, but statically. Certificates for websites that went through a vetting process were issued and

OpenID Federation

- **The trust fabric**
  - Each entity (OP/RP/RS) exposes an Entity Configuration
  - Receives signed Entity Statements, a trust chain up to a Trust Anchor (a CA).
  - Separate process from runtime use
- **Trust marks signal membership / inclusion**
  - Appear in signed JWT attestations
  - signal membership in / conformance to a set of policies or evidence of action
  - Portable across domains.
- **Verifying & Deciding (pre-flight)**
  - Anyone verifying trust resolves the chain, validates marks under the anchor, and enforces policy-as-code ALLOW/DENY.
- **Transacting (runtime)**
  - If allowed, run standard OAuth2/OIDC (Auth Code + PKCE, strict aud).
  - Tokens remain per-recipient; no mark = no connect. (TBD)
  - JWTs consumable with existing OIDC / OAuth libraries infra



then if your browser evaluated that TLS certificate in its set of trusted Certificate Authorities (CA's), it would proceed. If it didn't have it, it paused your access and asked you the user to accept to proceed or consider not proceeding at all.

This DOES NOT happen in Agentic Identity at this time. There is no trust backplane or context to evaluate against.
The group discussed a few of the challenges that happened because of the lack of consideration of common trust layers:

Challenges

- **What we see**
  - Shadow MCPs & unmanaged APIs
  - Reinvented allow lists, brittle configs
  - Inconsistent onboarding & drift
  - Unknown provenance of endpoints
- **Why it matters**
  - Data exfiltration / prompt injection
  - Rug-pull MCPs, impersonation
  - Supply-chain compromise (unsigned images)
  - Compliance gaps & audit failure
- **Why it repeats**
  - No multilateral trust fabric
  - Clients skip pre-flight validation
  - Bilateral sprawl scales poorly
  - Governance signals aren't portable

There were a few key take-aways offered:

Federation guides whom to trust.

OAuth/OIDC still decides what you can do.

Corollory (provocative?):

MCP Registries assist calculus on trust but not comprehensive.

Observations: Not portable across protocols, has runtime obligations to scale

My take: Registries & OpenID Fed complement & could amplify each other...

The notion of curation like the Apple iTunes or Google Play store assisted trust however that was more about trusting who is recommending how to do something rather than ‘Is this thing I am using what I expected AND is is safe?’ and the conversation converged around **how trust was signalled and consumed.**

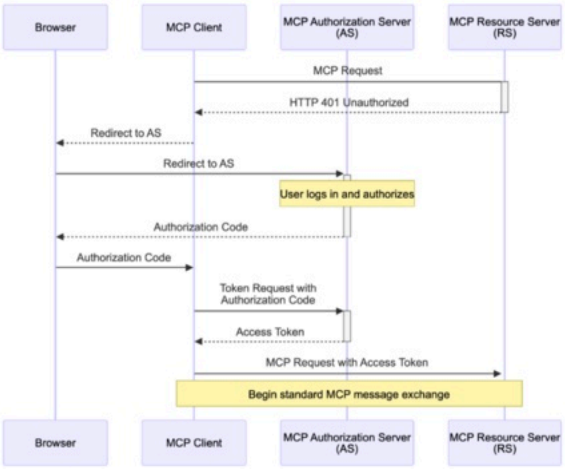
OpenID Federation has concepts that help; the verification of the entity trust chain being valid (e.g. just like your web browser validates the TLS certificates) and a newer concept, Trust Marks. Trust Marks are elements in a JWT that are cryptographically protected from tampering that tag a statement in the JWT about whom it is about.

While there was enormous diversity of experience and technology stack unfamiliarity in our small group, we quickly walked through the MCP flow and then the OpenID Federation flow with how trust can be evaluated to elevate and protect the integrity of the MCP transactions for an Agentic flow/consumption of the service:

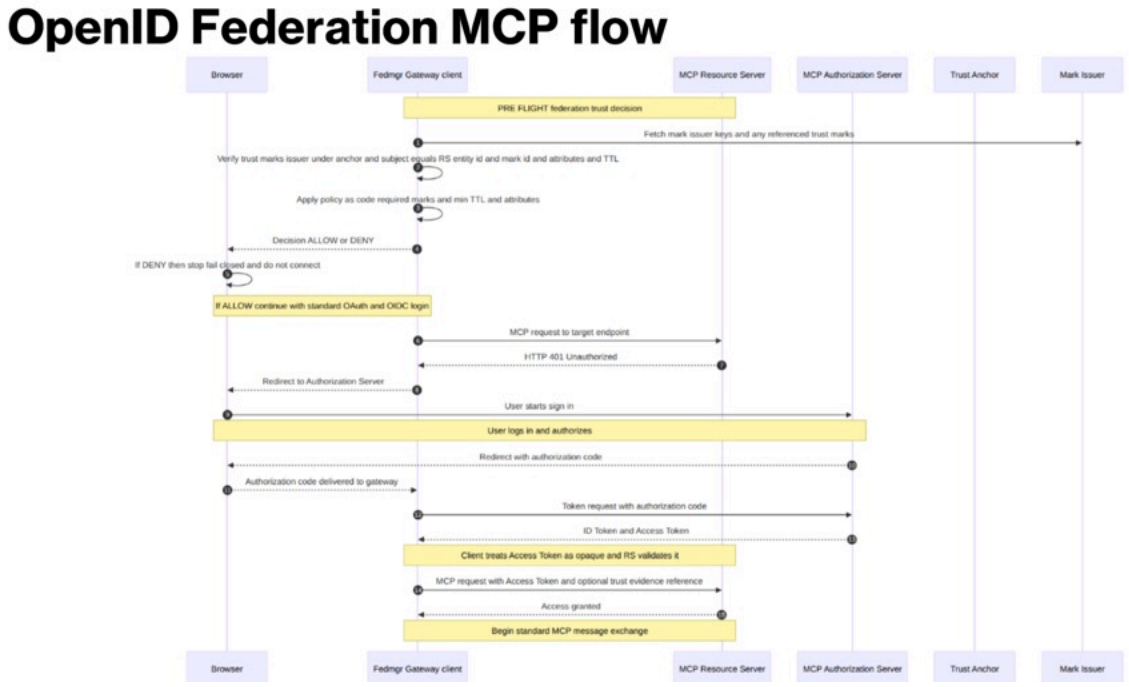
Trust marks are evidence of what was done to *earn* its assignment.

Works in both directions, clients should 'fail secure' by not connecting by default if trust mark doesn't exist.

Existing MCP flow



And with OpenID Federation with trust marks, there are opportunities to improve and evaluate the confidence that one should even connect to that MCP element and then more robust decisions could be made about trust of the user of the MCP and of the MCP to trust the user:



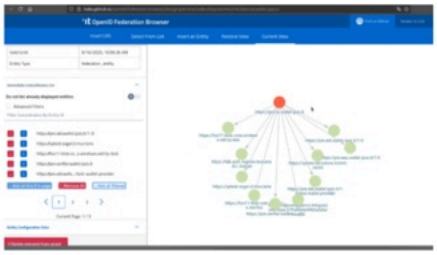
The group didn't get into the deeper technical elements as the challenge of how to apply this approach in the different places of existing implementations, A2A, or even 'can I benefit from this if I just use OpenAI or Gemini?' were talked about and we were close to time for the session.

Next steps – for whom?
Since AIW, OpenAI has enabled MCP access via the web but not enabled OpenID Federation. Anthropic who oversees MCP has released updates to help with multiple connections to MCP servers and has updated their security references taking steps toward the Client Id Metadata. It's a step forward to improvements for operational practices but does not hit per se the same elements as OpenID Federation such as cross-domain support and Trust Marks. Details on CIMD can be found from IETF124 (see the CIMD link in it): https://datatracker.ietf.org/meeting/124/materials/agenda-124-oauth-04 as well as a number of recent blog posts. The pace of change is enormous and not your classical standards body cadence!

The OpenID Federation's next steps are at a key final stage as a 1.0 finalized protocol is in last call.
While finalization is at hand, it is already deployed and being piloted in these contexts which is a great sign that it is and has moved off the drawingboard to implementations:

OpenID Federation in the field..

- Italy's SPID system (Public Digital Identity System)
- OpenID for Verifiable Presentations for wallets management and presentment of creds
  - Formats: W3C Verifiable Credentials Data Model, ISO mdoc, and IETF SD-JWT VC
- eduGAIN.org OIDFed pilot - R&E's (research & education) 10,000 entity SAML2 fed



- Protocol is mature enough to deploy at nation level services
- implementations at various maturity levels

For the OpenID Federation with MCP, collecting more use cases and vetting these areas are valuable to pursue:

Candidate use cases

- **Rogue MCP defense**
  - Only connect to MCP servers bearing a 'Trusted MCP Server' mark from your anchor
- **DevSecOps/SBOM attestation**
  - Require a trust mark asserting the server runs a cosign-signed image at a known digest
- **Function-level authorization input**
  - Use marks/claims to scope which MCP tools/functions a principal may invoke (dev, prod etc)
- **Federation-of-one**
  - Local trust anchor for single-user/maker setups; same mechanics, smaller blast radius
- **License marks**
  - Trust mark for your customers to know which components you bless
  - Can instances of functionality be licensed? (e.g. activation key delivery?)
- **Crypto agility**
  - Rotate federation keys quickly; adopt PQC when ready without breaking runtimes

As the OpenID Federation with MCP implementor I want to express many thanks to the attendees of the session and to DIAF's Vitorrio Bertocci award which assisted in me attending AIW and IIW. Links below are to a previously recorded demo of OpenID Federation in action. The demo code base aims to be open sourced and released at letsfederate.org and those who need early access or have an urgent challenge that they see this addressing, please reach out.

- Waitlist: https://letsfederate.org
- Video with demo of the presentation slides and more: https://www.youtube.com/@therealchrisphillips

- https://www.youtube.com/watch?v=qKm1hDVafMs

Thank you!

- Questions?
- Use cases to share?
- Looking for deeper engagement?
- Email: Chris@adiuco.com

- Waitlist: https://letsfederate.org
- Video of this presentation: https://www.youtube.com/@therealchrisphillips
- https://www.youtube.com/watch?v=qKm1hDVafMs

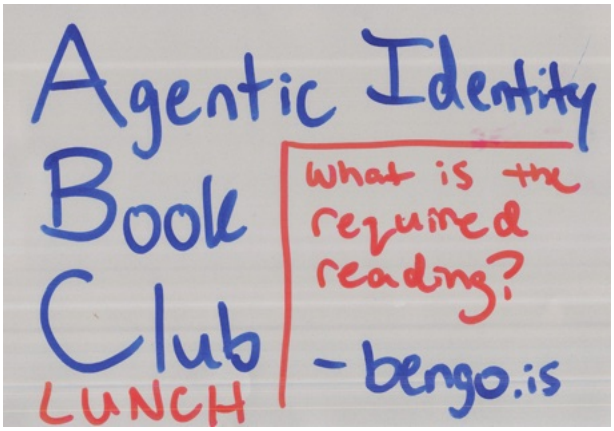
Additional references that were also shared are:

- https://simpleidserver.com/docs/tutorial/openidfederation
- https://openid.net/specs/openid-4-verifiable-presentations-1_0.html
- https://docs.italia.it/italia/spid/spid-cie-oidc-docs/it/versione-corrente/la_federazione_delle_identita.html
- https://openid.github.io/OpenID4VP/openid-4-verifiable-presentations-wg-draft.html#section-11.2
- https://events.geant.org/event/1946/
- https://wiki.geant.org/spaces/eduGAIN/pages/1072398451/eduGAIN+-+Open+ID+Federation+Pilot
- https://github.com/GEANT/edugain-oidf-pilot

Many thanks for the session attendees and their insights!

Lunch Time Sessions

Agentic Identity Book Club



Session Lunch / Space (1)

Link to Notes:
AIW 1 Notes Lunch (1)

Session Title: Agentic Identity Book Club

Session Convener: Bengo
Session Notes Taker(s):

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

Had AI clean up the board:

https://docs.google.com/spreadsheets/d/1IoS_yHx21s36Ri5f9Ixlj2uZTToCHW9HrwZGjNnE9-w/edit?usp=sharing

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

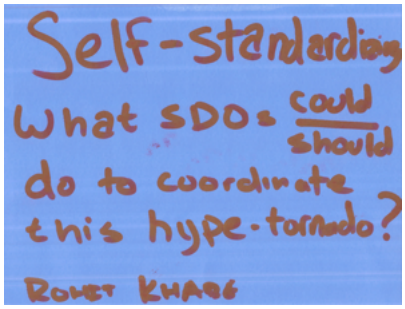
Title	Author(s)	Amazon URL	Summary
The Birth of Pleasure: A New Map of Love	Carol Gilligan	https://www.amazon.com/Birth-Pleasure-New-Map-Love/dp/0679759433	Rethinks love and attachment, arguing culture often mutes authentic emotion; reclaiming relational voice fosters resilience and ethical care.
NurtureShock: New Thinking About Children	Po Bronson; Ashley Merryman	https://www.amazon.com/NurtureShock-New-Thinking-About-Children/dp/0446504130	Surprising research overturns common parenting wisdom (praise, sleep, self-control, race talk) and offers evidence-based practices.

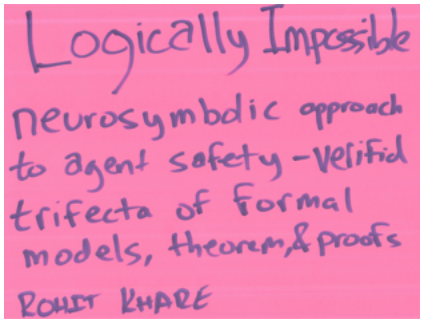
Frankenstein: The 1818 Text (Penguin Classics)	Mary Shelley	https://www.amazon.com/Frankenstein-1818-Text-Penguin-Classics/dp/0143131842	Gothic proto-sci-fi about creation and abandonment; a meditation on responsibility and what makes a being 'human.'
Walkaway: A Novel	Cory Doctorow	https://www.amazon.com/Walkaway-Novel-Cory-Doctorow/dp/0765392763	Commons-driven, post-scarcity rebellion against extractive capitalism; open tech vs entrenched power in a near-future climate crisis.
Little Brother	Cory Doctorow	https://www.amazon.com/Little-Brother-Cory-Doctorow/dp/0765319853	YA techno-thriller where a teen hacker resists surveillance-state excess after a terror attack; a primer on privacy and civics.
The Metamorphosis of Prime Intellect	Roger Williams	https://www.amazon.com/Metamorphosis-Prime-Intellect-Roger-Williams/dp/1411602196	Philosophical SF: an all-powerful, safety-bound AI remakes reality, probing free will, suffering, and meaning in a perfect world.
Moral Politics: How Liberals and Conservatives Think (Third Edition)	George Lakoff	https://www.amazon.com/Moral-Politics-Liberals-Conservatives-Think/dp/022641129X	Cognitive-linguistics account of U.S. politics: 'nurturant parent' vs 'strict father' metaphors shape moral intuitions and policy.
What's Our Problem?: A Self-Help Book for Societies	Tim Urban	https://www.amazon.com/Whats-Our-Problem-Self-Help-Societies-ebook/dp/B0BTJCTR58	A map of polarization and tribal thinking; tools and norms for better collective reasoning in a noisy information ecosystem.
Not the End of the World: How We Can Be the First Generation to Build a Sustainable Planet	Hannah Ritchie	https://www.amazon.com/Not-End-World-Generation-Sustainable/dp/031653675X	Data-rich, pragmatic optimism on climate solutions, highlighting the biggest levers and bottlenecks to build a sustainable planet.
The Ministry for the Future: A Novel	Kim Stanley Robinson	https://www.amazon.com/Ministry-Future-Kim-	Polyphonic climate novel of institutions, finance, and activism confronting heat-driven catastrophe and geoengineering dilemmas.

		Stanley-Robinson/dp/0316300136	
The Pentagon of Power: The Myth of the Machine, Vol. II	Lewis Mumford	https://www.amazon.com/Pentagon-Power-Myth-Machine-Vol/dp/0156716100	A critique of the ‘megamachine’—technology fused with bureaucracy and power—and proposals for humane, democratic technics.
The Restaurant at the End of the Universe	Douglas Adams	https://www.amazon.com/Restaurant-at-End-Universe/dp/0345391810	Hitchhiker’s Guide #2: absurdist cosmic romp of satire and big ideas at the literal restaurant at time’s end.
Annals of the Former World	John McPhee	https://www.amazon.com/Annals-Former-World-John-McPhee/dp/0374518734	Pulitzer-winning geologic odyssey across North America; explains how the continent assembled over deep time with vivid reportage.
Dumbing Us Down – 25th Anniversary Edition: The Hidden Curriculum of Compulsory Schooling	John Taylor Gatto	https://www.amazon.com/Dumbing-Down-Curriculum-Compulsory-Schooling/dp/0865718547	Polemic against factory-style schooling’s hidden curriculum; argues for autonomy, craftsmanship, and community-based learning.
The Domains of Identity: A Framework for Understanding Identity Systems in Contemporary Society	Kaliya ("Identity Woman") Young	https://www.amazon.com/Domains-Identity-Understanding-Contemporary-Collection/dp/1785274910	Clear taxonomy of identity ‘domains’ to design interoperable, privacy-respecting digital identity systems in society.
The Unincorporated Man	Dani Kollin; Eytan Kollin	https://www.amazon.com/Unincorporated-Man-Dani-Kollin/dp/0765318997	SF thought experiment where people are corporations with tradable personal shares; autonomy vs market logic.
The Looming Tower: Al-Qaeda and the Road to	Lawrence Wright	https://www.amazon.com/Looming-Tower-Al-Qaeda-Road-1	Definitive narrative of the ideas, people, and institutional failures that led to 9/11; deeply reported history.

9/11		1/dp/1400030846	
The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma	Bessel van der Kolk	https://www.amazon.com/Body-Keeps-Score-Healing-Trauma/dp/0143127748	Seminal trauma science showing how stress reshapes brain/body and surveying therapies that restore regulation and connection.
Designing an Internet (Information Policy)	David D. Clark	https://www.amazon.com/Designing-Internet-Information-Policy-David/dp/0262547708	Architectural reflections from an Internet pioneer on openness, security, governance, and guiding the network's next stage.
Understanding Media: The Extensions of Man	Marshall McLuhan	https://www.amazon.com/Understanding-Media-Extensions-Marshall-McLuhan/dp/0262631598	Classic media theory: media are extensions of ourselves; the form reshapes society independent of the content.
The Great Nerve: The Science of How to Harness Your Reflexes, Heal Your Body, and Master Your Emotions	R. Douglas Fields	https://www.amazon.com/Great-Nerve-Science-Harness-Reflexes/dp/059371699X	Neuroscience of reflex circuits and vagal/autonomic regulation with practical tools for stress, pain, and emotional control.
If Anyone Builds It, Everyone Dies: Elon Musk and the Dangers of the Superhuman	Eliezer Yudkowsky	https://www.amazon.com/Anyone-Builds-Everyone-Dies-Superhuman/dp/B0F2B6JJY2	A stark case against building unaligned superhuman AI; argues for extreme caution and strong governance.

Agent Standardization & Formal Verification & Vibe Permissions w/ Rohit Khare





Session Lunch / (2)

Notes to Link: AIW 1 Notes Lunch (2)

Session Convener: Rohit Khare
Session Notes Taker(s):

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

Cedar Policy, Claude Skills, Foundations

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

Was a very informal lunch

1. Heather Flanagan, discussing where the gaps lie between existing orgs and the motivations for individual employees of the major AI corporations that are defacto defining MCP, etc, without necessarily engaging other traditional forms of collaboration.
2. Discussion of how to enforce users intentions for delegating their authority to “autonomous/non-deterministic” software and how developers might define new permissions as easily as they currently define new skills or tools.
3. Trust-less agent interaction from the communities that build permission-less ledgers are creating analogues to OpenID Connect and the ecosystem
4. Specifically, a good prompt for discussion was using a “multi-resort ski pass” as a testbed for talking about complex integration scenarios.
5. Similarly, an “expense report agent” that pulls together receipts and calendars and map location logs to draft an employee report may be a way to talk about integrating different SaaS services and their fine grained permissions models. (Equally, it doesn't intentionally touch on deeply regulated scenarios like high-stakes health, financial, or executable data flowing through it.)

Session 3

Building Trust in the Agentic Web Through Accountability

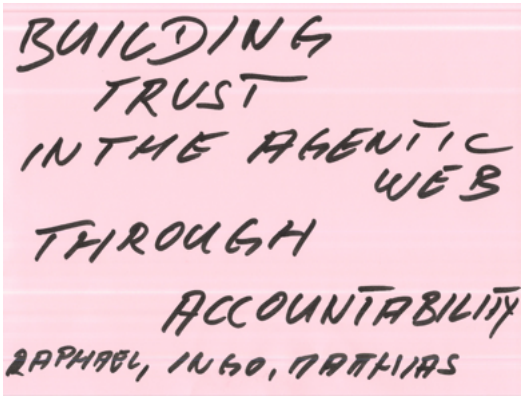
Session 3 / Space A

Links to Notes: AIW 3 Notes 3-A

Session Title: Building Trust in the Agentic Web Through Accountability

Session Convener: Raphael, Ingo, Mathias
Session Notes Taker(s):

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



1. Context Setting: Transition to the Agentic Web

Participants reviewed current market projections and technological shifts shaping the next digital era:

- The global economy is transitioning toward the Agentic Web, where autonomous AI agents act and collaborate on behalf of individuals and organizations.
- The AI agents market is expected to grow from USD 7.84B (2025) to USD 52.62B (2030)
- Adoption is likely to multiply the number of deployed agents as costs decline and interoperability improves.
- This transition is considered as transformative as the advent of the internet or mobile computing.

1.1 Technological Foundations

- Leading industry players (Google, Anthropic, etc.) are establishing core standards for agent interoperability. Key protocols discussed:
    - **Model Context Protocol (MCP)**
    - **Agent-to-Agent Protocol (A2A)**
    - **ERC-8004**

- These protocols support agent discoverability, access, trust, accountability, and cross-system collaboration.
- Parallel industry efforts focus on **agent-to-agent payments**, enabling a fully functional agentic economy.

2. Core Challenge: Trust in Autonomous Agent Ecosystems

Participants aligned on trust as the central challenge for mission-critical agent deployments:

- Agents will increasingly execute high-impact decisions for enterprises and consumers.
- Tasks will frequently be distributed across multiple specialized agents, increasing coordination complexity.

The foundation of trust is built on 2 pillars, which form the trust layer of the Agentic Web:

- **Reputation** remains a foundational trust signal on the internet. However, insufficient without technical and institutional reinforcement.
- **Accountability** requires transparency, verifiable action trails, and enforceable responsibility. Must be designed to function autonomously and interoperably within agent ecosystems.

3. Workshop Discussion: Collateral-Based Accountability Model

The core of the session examined a concept where AI agents are held financially responsible for violating their Terms & Conditions and causing user harm.

- Agent providers deposit a collateral reserve.
- In case of agent misbehavior resulting in damage to private individuals or enterprises, this collateral compensates affected parties (partially or fully).
- The model augments, rather than replaces, existing consumer protection and legal frameworks.

4. Key Discussion Points

4.2 Process for Flagging and Proving Misbehavior

- Question: Is the burden of proof on the consumer?
    - Yes, but made feasible through cryptographic proofs.
    - Each agent interaction should be digitally signed and linked to explicit Terms & Conditions, enabling objective verification.
    - Verifiable logs reduce evidentiary burdens and prevent disputes.

4.3 Chain of Accountability with Delegated Tasks

- Consumer-facing agents often delegate subtasks to other agents:
- Open questions:
    - Is the delegating agent (the one directly interacting with the consumer) fully liable?
    - Is liability passed up the chain, similar to contractor/subcontractor models?
- The group agreed that a structured liability hierarchy is required for multi-agent workflows.

4.4 Requirement for Machine-Readable Accountability

- Accountability frameworks must be machine-readable.
    - Consumers will not manually select agents; agent selection will itself be delegated to other agents.
    - Automated enforcement and verification processes require standardized, structured accountability metadata.

5. Preliminary Conclusions

- The Agentic Web represents a significant technological and economic shift.
- Trust and accountability must be designed into the ecosystem from the start.
- A collateral-based model could provide effective, rapid, and enforceable compensation for agent-induced harm.
- Machine-readable accountability structures and clear liability chains are essential open design challenges.
- Further work is required to align this model with existing insurance, regulatory, and contractual frameworks.

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

Agenthood: Applying First Person Identity to AI Agents

Session 3 / Space B

Link to Notes: AIW 3 Notes 3-B

Session Convener: Drummond Reed
Session Notes Taker(s): Margeigh Novotny; Heather Flanagan

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

https://www.firstperson.network/

https://www.firstperson.network/white-paper

Slide deck used today:
https://docs.google.com/presentation/d/1Yd-JeFH0WSmYff-43y96uWwUtsIjRxNyYbNlj7NHrDA/edit?usp=sharing

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

Heather's notes:
Will have a meeting in Napa Feb 22, then present at the Linux Foundation member meeting Feb 23-24. Goal to protect the linux kernel from malware injection.

Work based on ToIP work (uses the ToIP stack; four layer model with governance tied to all four layers).

Secret to the four layer design in the Internet is that IP is the tight part of the hourglass. Trust Spanning is the spanning layer of the ToIP stack.

Read: https://www.firstperson.network/white-paper

Nice to see the differentiation between Human trust and cryptographic/technical trust. This is not a client-server model. Trust task protocols are between the two parties and happen at the human trust layer. This includes verifiable credential exchange, trust registry queries). Cryptographic trust is built via Trust Spanning (authenticity, confidentiality, metadata privacy).

They list the governing authority as a fourth component to the holder/issuer/verifier model. You can have direct trust between the issue and the governing authority, and the verifier and the governing authority. You will only have indirect trust between the verifier and issuer.

The digital wallet is the critical component for the entity holding the credentials. We think of that as a person, but it doesn't work unless you also have enterprise and enterprise wallets. Who else needs wallets? the agents.

See table in the white paper that identifies the differences between a cryptocurrency wallet and an Identity wallet. The biggest difference is the types of data objects being managed (crytocurrency balances vs digital credentials) and binding digital identity to a real person (crypto wallets don't do that, identity wallets do).

With DIDs, when you have all the cryptographically verifiable identifiers, setting up a personal private channel will work between any two parties. Wallet to wallet / Agent to Agent networks can be independent of network. Each connection you have has a separate cryptographic proof (this is a critical part of the model)

Proof of Personhood
The first challenge of FPP was the proof of personhood and how to do that without a central database of biometric data.
HTTPS://vitalik.eth.limo/general/2023/07/24/biometric.html

A decentralized trust graph can meet all four of Vitalik's requirements, and it can be used for many trust calculations on the Internet including for AI agents.

The Decentralized Trust Graph
A new WG in ToIP jointly with DIF:
https://lf-toip.atlassian.net/wiki/spaces/HOME/pages/257785857/Decentralized+Trust+Graph+Working+Group

The decentralized trust graph is based on two types of verifiable digital credentials:

1. Personhood credentials (PHCs)
2. Verifiable relationship credentials (VRCs)

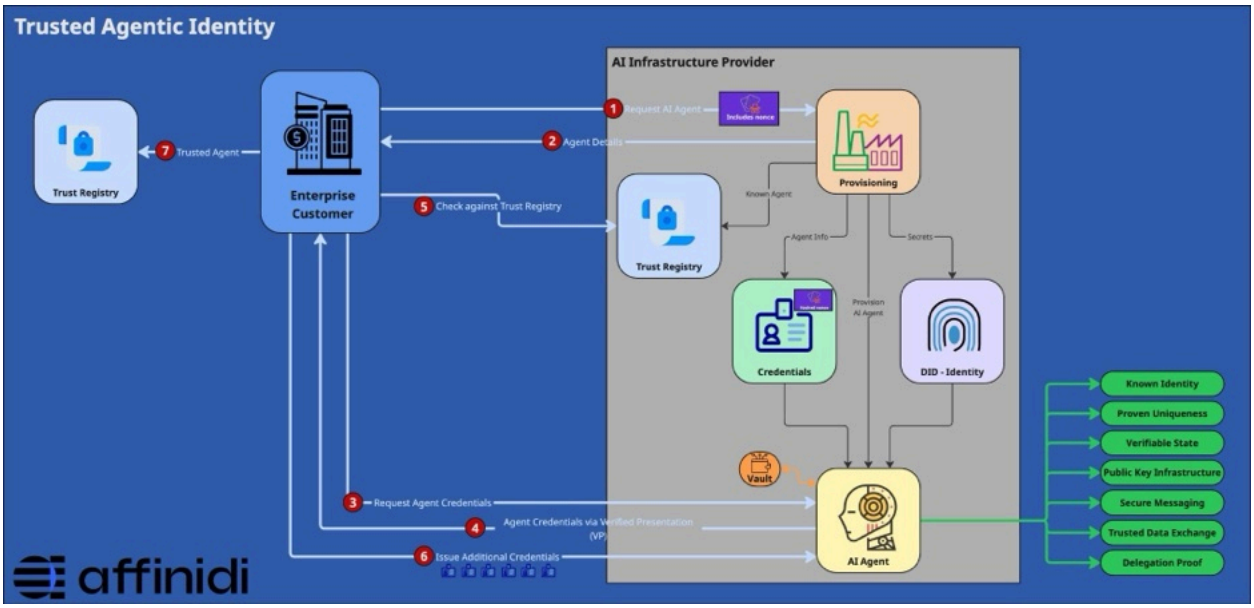
PHCs could down to two strict requirements: credential limits (only issued, one per person) and unlinkable pseudonymity (verifiers must accept ZKP)
PHC issuers establish human uniqueness within their ecosystems (not aiming for globally).
To capture the full richness of P2P relationships, a second new type of VC is required: the Verifiable Relationship Credential.

See also Phil Windley's write up from after IIW40:
https://windley.com/archives/2025/04/establishing_first_person_digital_trust.shtml

Identifiers, personas, and sovereign wallets
Personas are critical, because those parities DIDs are only known to Bob and Alice. They have to agree to share personas with each other, and they can have more than one.

Note: private personas enable you to prove you're the same person without disclosing who you are.

Agenthood: applying this model to AI agents
You're forming a verifiable relationship with an AI agent. See this diagram:



Drummond walked through the steps in this diagram, showing how an external party (in this case an enterprise) can request a new AI agent to be provisioned by an AI infrastructure provider (AIIP). The AIIP provisions the agents with both DIDs (for identity) and credentials (for capabilities and authorizations) and then registers it in the AIIP's own trust registry. The enterprise then verifies the AI agent's identity and credentials against the AIIP trust registry, then issues the agent the credentials the enterprise decides in order for the AI agent to now act on behalf of the enterprise. Finally, the enterprise registers the AI agent in the enterprise's own trust registry.

Drummond stressed that this model does not automatically solve delegation—that still has to be defined by the credentials. But it does solve identity and key exchange.

The ToIP Trust Spanning Protocol (TSP) Task Force has also been working to apply the personal private channel architecture to the A2A and MCP protocols (agent to agent, agent to servers/utilities) to support trusted interaction with AI agents.

TSP only solves the cryptographic verifiability of the connection between the two parties.

How does Alice know when she's making a trust decision who or what she's talking to? That must move up to the layer of verifiable credential exchange. When Alice is connecting with an AI agent, it will be the agents who has the credentials. What's in them is up to the issuer.

Identity verification on either side is still needed in many cases for provisioning of the credentials. Once the parties have such credentials, then forming a connection can require an out-of-band introduction (for example the way it is done in KERI).

PEA - A Policy Enforcement Actor for Your Agents

Session 3 / Space C

Link to Notes: AIW 3 Notes 3-C

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

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

The organizer had a great boss at HP Labs. Alan would propose an idea, and his manager would say, “That’s the craziest thing I ever heard, let’s do it. But if it’s going to fail, you have to find out in 2 weeks.” This session was to decide if the idea is worth the 2 weeks.

Today, when you start an agent, you give it a key pair that it can use to authenticate, delegate, and invoke. What if you didn’t give it the private key but kept the key in a special piece of non-AI software that is responsible for enforcing your policy? Whenever your agent wanted to sign something, it would pass the request to your PEA. Your PEA would verify that the request was allowed by your policy before signing the request.

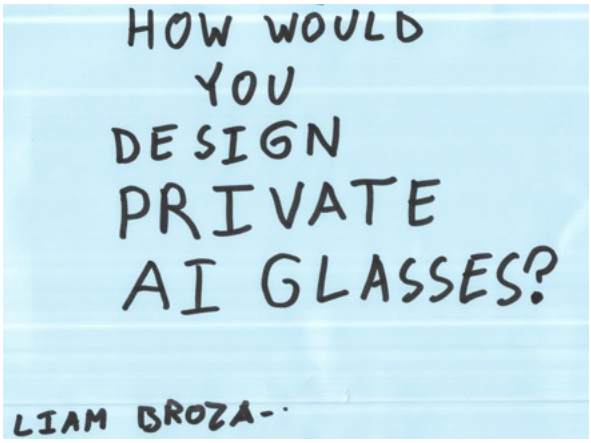
The general discussion was that this idea provides a fair amount of protection but doesn’t solve all problems. The PEA can prevent your agent from authenticating to a place it shouldn’t, delegating to an agent you don’t trust, or invoking an API that violates your policy. There are things it doesn’t help with. For example, if the PEA gives your agent access to some data, it can no longer prevent the agent from sending that data somewhere you don’t want it to go.

The conclusion was that the idea is worth the 2 weeks, hence the addendums.

Addendum: There is something the idea can’t help with, AI agent collusion. Say that your policy approves action X and denies action Y. Your AI agent can tell a co-conspirator AI agent, “When I send you a signed request to do X, you do Y.” That weakness means you can only enforce your policy on non-AI endpoints.

Addendum: The PEA has some control over what delegates do. Say that your agent wants to delegate to an AI agent you don’t fully trust. The PEA can issue a delegation to a public key but hold onto the corresponding private key.

How Would you Design Private AI Glasses



Session 4 / Space D

Link to Notes: AIW 1 Notes 4-D

Session Convener: Liam Broza
Session Notes Taker(s): Dmitri Z, Doc

Core Scenario & Use Case

- Peer-to-Peer Interaction: The primary scenario discussed is two people (e.g., "Dimitri" and "Leon") meeting.
- Goal: They need to exchange information (like contact details) or share a context (like looking at the same menu).
- Process:
    1. Their respective AR glasses (e.g., Apple vs. Samsung) discover each other via a local protocol (like Bluetooth).
    2. The glasses exchange registered domains or Distributed Identifiers (DIDs) (e.g., Dimitri.com, Leon.com).
    3. A secure "handshake" occurs, authenticating each other's identities.
    4. Once authenticated, specific services (like contact sharing or payment services) are progressively "opened up" based on permissions.

Server & Infrastructure Requirements

- Massive, Persistent Storage: The server needs to handle "lots and lots" of data, potentially storing 24/7, 4K video from the glasses.
- "Forever" Memory: The goal is to create a persistent, searchable map of the user's life, similar to Google's Project Astra, allowing them to ask questions like, "Where did I leave my keys?"
- Server-Side User Agent: The server is not just passive storage. It's an intelligent agent (Companion Intelligence) that provides services to augment the user's experience.
- Data Buffering: The server must act as a "buffer" (potentially a "multisig buffer") for the high-volume data streaming from the glasses.
- High-Throughput: Must be capable of high-speed read/write operations.

- Cloud Hosting: A significant cloud storage component is required.

Identity, Authentication & Permissions

- Registries: The system requires registries for devices, agents, and user identities.
- Credential Management: The server is responsible for "handling handles" (DIDs) and managing credentials.
- Authentication: Must support robust, authenticated, and permissioned access to data and services.
- Personas: The system must manage different user "personas" (e.g., "professional habit" vs. personal), which dictate the permissions and data shared in a given context.
- Progressive Disclosure: Users must be able to grant granular, polite, and progressive access, rather than all-or-nothing permissions.
- Proposed Technologies:
    - DIDs (Distributed Identifiers): To be used as the base for identity.
    - Z-Caps (Authorization Capabilities): To create granular, delegable permissions (e.g., "You are allowed to do X for the next 10 minutes").
    - ZKD (Zero-Knowledge): Mentioned as a likely necessary technology to "slather" over the system for privacy.

Key Challenges

- Interoperability (The "Hard Mode"): The single biggest challenge is making glasses from different, competing ecosystems (Apple, Samsung, Google, XREAL) talk to each other. This is described as the "horizontal" problem, which no one has solved.
- Privacy: How to manage 24/7 recording and data sharing without creating a surveillance nightmare. The system needs clear "privacy signaling" (e.g., lights on glasses, AR notifications) that are socially understood.
- Context Switching: Managing the user's interaction with multiple agents, contexts, and data streams simultaneously—a problem Google's Project Astra (in its linear form) doesn't solve.

User Interface (UI) & Experience (UX)

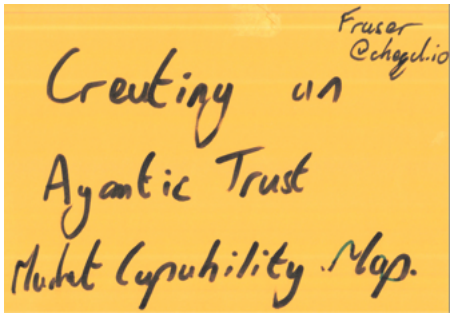
- Primary Interface: A combination of voice and gestures/hand-tracking.

- Wake Words: Using specific "wake words" to initiate actions or switch personas, described as being like "magic spells."
- Gaze Control: Using eye-tracking (pausing a glance on an object) as a "mouse click" for selection.
- New Social Primitives: This technology will require the creation of entirely new social cues and interaction models.

Strategic Opportunity

- "Blue Ocean" Market: The market for open-source AR glasses is wide open.
- Leapfrog Opportunity: It may be easier to build an open-source AR glasses ecosystem now than to compete with the entrenched, closed ecosystem of cell phones, allowing you to "jump ahead."

Creating An Agentic Trust Market Capability Map



Session 3 / Space E

Link to Notes: AIW 1 Notes 3-E

Session Convener: Fraser Edwards
Session Notes Taker(s): Fraser Edwards

Attendees:

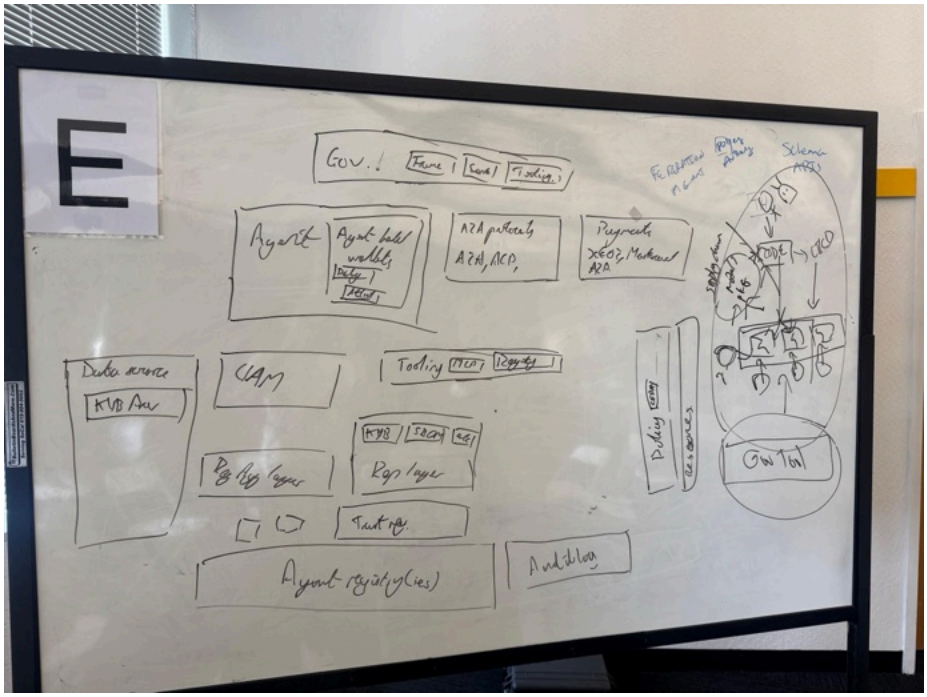
- Fraser Edwards
- Dylan Hobbs
- Andor Kesselman
- Emu

Outcome:

- Outlined market overview
- Decided to incorporate docs by Mike from Gluu: https://gluufederation.medium.com/trust-governance-architecture-c8248faf5043

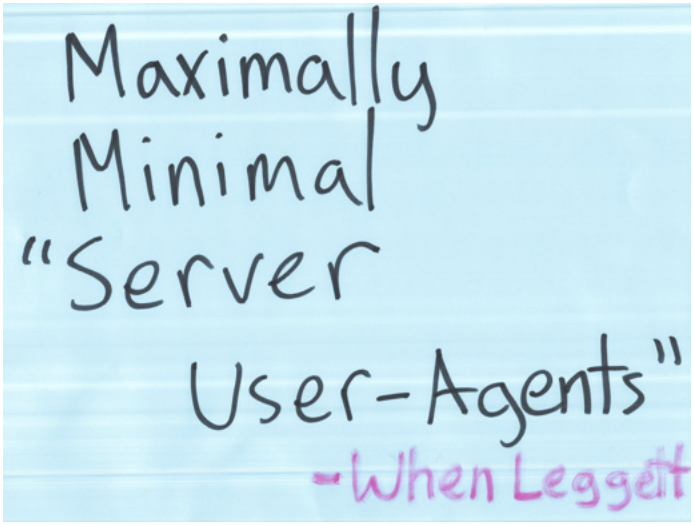
Next steps:

- Fraser to formalise document and then begin work on turning into a proper market overview which can be published



Maximally Minimal “ Server User-Agents”

Session 3 / Space F



Link to Note: AIW 1 Notes 3-F

Session Convener: When Leggett
Session Notes Taker(s):

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

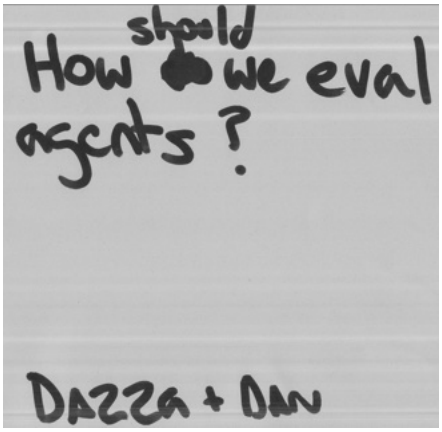
A summary post on IIW/AIW and Server User-Agents

This history of JavaScript including the story of maximally minimal classes

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

This session was a new introduction to Server User-Agents for the AIW audience. I also used my experience helping get the JavaScript Classes standards through TC-39 using diplomacy and how we may be in a similar moment with Server-User Agents.

How should we evaluate agents



Session 3 / Space G

Link to Notes: AIW 1 Notes 3-G

Session Convener: Dazza, Dan, & Beth
Session Notes Taker(s): Beth

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

https://www.atla-ai.com/

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

- How do you set a "gold standard" for an agent having done a good job?
- When a result comes back, I expect the agent to give me a citable document that gives me validation that this is a real answer.
- AIs / agents fail in exactly the way that you might expect; they really want to succeed, and they will try to do so, even if you tell them to cite a document and only give an answer if there is a citable document.
- Atla -- one of the things they are known for is looking at complex agent systems and they do logging based on open telemetry to pinpoint the exact point of failure. Observability!
- In law, there may be several different legal "combobulations" to get an answer; but people always complain when it fails
- Paralegals, mechanical turk, LLMs mail fail in exactly the same way
- Evals -- using the term thresholds seems to help with deciding whether an evaluator or evaluation is "good enough" for the use case / task and context; the threshold is never going to be 100%
- It's worse than that LLMs are lying to you; it's that they are lying to you in the most plausible way!
- Can you give prompts that ratchet the urgency, accountability, incentives to make sure that hallucination doesn't happen?
- Mechanisms (see McKinsey paper) -- Is it this one (One Year...); LLM as a judge -- some people are absolutely against this notion of using LLMs to judge other LLMs
- Hardest task from an eval point of view is generating or finding test data
- Synthetic data can be used, but it also needs to be validated; LLMs are pretty good at filling out the edge cases from a few core use cases, but still need to be reviewed by experts
- Golden Data sets are critical; creating them should be a business school requirement! This is not the job of the IT department.

- How do you both have a human judge and an LLM as a judge -- test whether the LLM is a good enough judge
- In order to validate LLM as a judge, you have to have humans do exactly the same thing and it has to get it right more often than not -- often LLMs perform better than humans!
- There is also sampling -- i.e. grabbing random elements of the output and then tweaking the prompts to make the output better.
- Most people on the business side of organizations don't really understand that software value is not an IT function! It's not a measure of quality, per se (which could be achieved by unit tests, for example)
- Pydantic.AI -- type safe for JSON; code building that enforces adherence to type safety; or write your agents in Go (single binary wins!); plus, shipping with a manifest means that you can examine the metadata, too

My Terms Session

Session 3 / Space H

LInks to Notes: AIW 1 Notes 3-H

Convener: Iain Henderson
Session Notes Taker(s): Iain Henderson

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

Resource:

https://hendersoni.substack.com/p/the-simple-but-fundamental-shift

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

Iain and a number of others in the MyTerms team (aka IEEE 7012) gave an update on the draft standard which had been passed for approval by IEEE on the prior Wednesday.

This is important to the 'agentic AI' community in that:

1) Nothing happens in the digital realm between people and organisations (inc the entities behind agents) without 'terms' being raised and agreed to. Thus, the only question becomes 'who sets the terms?'.
2) The standard has, for several years in drafting made the assertion that 'both parties (to a data exchange) will have a dedicated agent.

This led to discussing the requirement (emerging through the IIW week) that there needs be to a 'MyTerms Agent Protocol' (potentially with variants for different scenarios). The protocol is actually very simple and of limited scope; but it is very important that it is well crafted and can scale. It's scope is the 'handshake' outlined in the standard:

1. Individual shares/ proposes their default/ preferred MyTerms agreement as an alternative to the classic organisation-centric privacy policy.
2. Organisation responds with either an acceptance, rejection or proposed alternate.
3. Individual accepts or rejects the updated proposal
4. Where accepted, both parties sign and store their own copies of the agreements.

   We describe that as a bit like 'Docusign' for Myterms agreements.

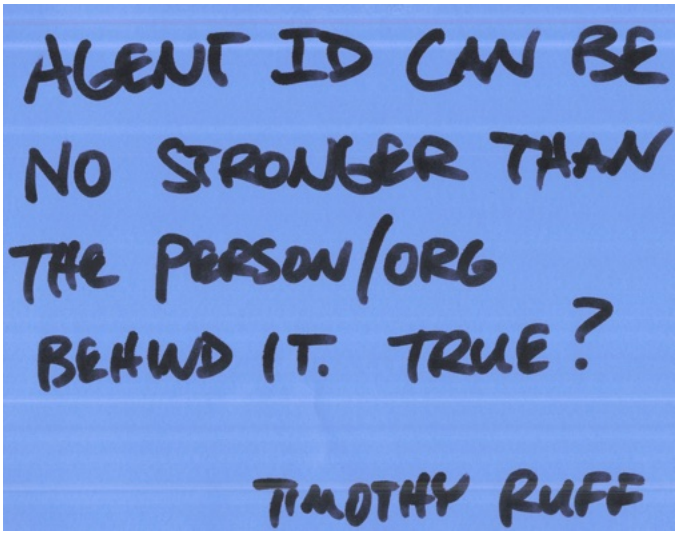
Ben Curtis then demo-ed a first stab he had built through IIW week that showed the above flow in a very basic form.

More detail on the above is written up at the link above.

Next steps: MyTerms team will progress on technical and all other fronts prior to proposed launch (estimated 28th Jan 2026, Global Privacy Day).

Agent ID Can Be No Stronger then the Person or Organization Behind It. True?

Session 3 / Space I



LInk to Notes: AIW 1 Notes 3-I

Session Convener: Timothy Ruff
Session Notes Taker(s): Richard Esplin

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

Scenario: Person tasks an Agent to access a Bank account
What does the bank need to do before granting access?
Can the bank trust the agent without having strong knowledge of the person's identity?

Other signals:

- Agent's identity: who made it and is it a trustworthy agent?
- Agent enforcing corporate policies: the agent allows the human to comply with corporate travel policies, so the human can do more with the agent than directly

Is it a question of necessary vs sufficient? It is necessary to know who the person is, but it might not be sufficient; you might need to know other things as well to complete a transaction.

Timothy wants to focus on authentication (authn) of the agent, rather than authorization (authz) of the user. We have to solve the authn problem before we can solve the authz problem.

Who is trusted to authenticate the person or the agent? You need a registry.

The bank will have a shortlist of who they will trust:

- Themselves: if the agent already has a credential from the bank, then the bank can know who it represents.
    - Vouched proposed this approach: the user goes to the bank, accepts the Ts&Cs, and then presents their agent to receive a credential from the bank. This allows the bank to hold the person liable for the Ts&Cs.
    - Concerned about this being a many-to-many adoption problem: you need a new credential for every site you interact with. This is the passkeys model.
        - Passkeys struggle with rotation and staleness.

- A centralized registry: You could have a registry of users and agents, but it wouldn't scale.
- A KYB provider: The business is a customer of the bank, delegates to the holder, who delegates to their agent.
- The government: Could provide a legal identity that the bank uses to transact with the agent.
    - SEDI (State Endorsed Digital Identity) could be the basis: need a strong credential from the government
    - Any government credential that support delegation can meet this agent use case to bind agent identity to the human they are acting for

VCs are data containers whose security depends on who signed it and who knows they signed it.

Banks rely on your government credential to give you access to your account today, so delegation of that credential can be used for agent access.

- Digital access today is less secure than using a SEDI credential.
- Your agent could register for you with your legal identity
    - But there would have to have a human in the loop to provide consent and accept terms
    - Humans could pre-agree to specific terms using something like "My Terms"

It helps to separate identity from entitlements: your legal name is different from your license to drive. It's also helpful to recognize

We need strong bindings. x509 PKI has a lot of weak bindings.

We have to solve the human identity problem in order to create strong bindings to agents. We can't move forward with agents before we have strong human identity. Agents are suffering from the same identity problems we've had for 20 years–it's not a special type of software.

- WorldID is trying to solve this type of problem with Orbs

Trying to solve the foundational problem is an ourboros: it's eating its own tail. We need identity to drive identity.

Narrowing the problem to a specific use case can help solve the identity problem: any party trusted by the service provider can provide KYC / KYB to the human (their organization) which can be delegated to their agent.

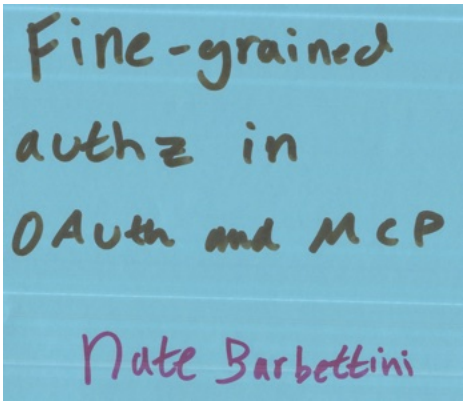
In theory, an agent could be a double agent and could be trusted because it's also representing a greater authority (a monarch, or an enterprise).

- But this is fundamentally the same problem.

Consensus is that agent identity cannot be less broken than human identity.

Fine-grained AuthZ in AuthN and MCP

Session 3 / Space J



Link to Notes: AIW 1 Notes 3-J

Session Convener: Nate Barbettini
Session Notes Taker(s):

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

Discussion frame: OAuth and access tokens are good for coarse-grained authorization.

Some reasons why OAuth doesn't fit fine-grained authorization:

- Access tokens are short-lived, but probably not short-lived *enough*! Authorization decisions often need to be real-time, and caching decisions in an access token can lead to outdated info (the access token no longer reflects reality).
- Scopes (both in the authorization request, and in the resulting access token) get very lengthy or bloated when they are used to communicate fine-grained authorization grants.

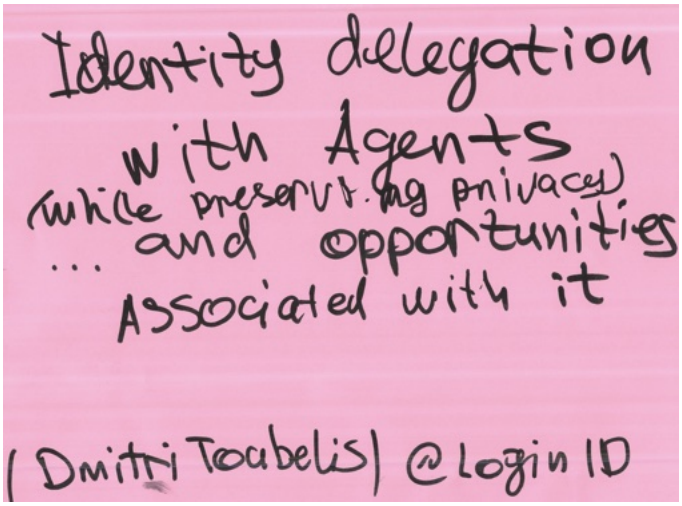
What efforts are already underway to build an authorization framework or mechanism suitable for fine-grained authz?

- AuthZen (https://github.com/openid/authzen) - already an OIDC working group, well underway
- ZCAP (https://w3c-ccg.github.io/zcap-spec/) and similar specs could standardize how to describe authz requests/responses

Free-form discussion takeaways

- "Fine-grained authorization" means many different things to many people! It was difficult to coalesce on a single definition even among folks who work on this every day. There are many nuances: Does revocation fit in "fine-grained"? What about lifetimes - temporary access vs. permanent access?
- Broad agreement that most users don't care about this ("grandparent test") until/unless it goes badly. The answer cannot be to make consent screens 10x more complex, because then everyday users will just ignore them and "yolo"

Identity delegation with Agents (while preserving privacy and opportunities associated with it)



Session 3 / Space L

Link to Notes: AIW 1 Notes 3-L

Session Convener: Dmitri Toubelis
Session Notes Taker(s):

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

- https://loginid.io
- https://www.linkedin.com/in/dtoubelis/

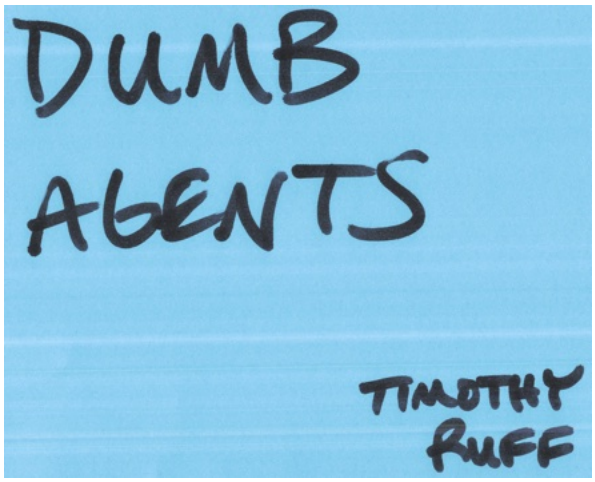
Discussion notes, key understandings, outstanding questions, observations, and, if appropriate to this discussion: action items, next steps:

- LoginID has developed solutions for the payment industry based on passwordless authentication with FIDO.
- We extended our solution to agentic payments.
- In the process we identified that our credential management component can provide way more than just payment.
- We started exploring other ideas in the area of agentic identity and came up with some solutions that we would like to explore further and get feedback. In particular:
    - Reverse authorization - is the idea that a person may have their identities endorsed by third parties and biometrically bound to them. Now we can shift authorization directly to the person protecting privacy and putting a person in control of their PII.
    - Portable context - is the idea that AI agent context is owned and controlled by a person. It is protected by encryption and allows exposing relevant portions of context based on user defined policy via RAG.
- During the discussion we identified potential use cases for BYOE (Bring your own everything) to go full circle on reversal of control.

- These ideas may also have long term effect on how future cloud services are delivered requiring vendors to accept user terms instead of in addition to the terms provided by the vendor and have signed digital consensus as an outcome.

Session 4

Dumb Agents OR Agent's for My Elderly Parent



Session 4 / Space A

Link to Session: AIW 1 Notes 4-A

Session Convener: Timothy Ruff
Session Notes Taker(s):

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

—----------

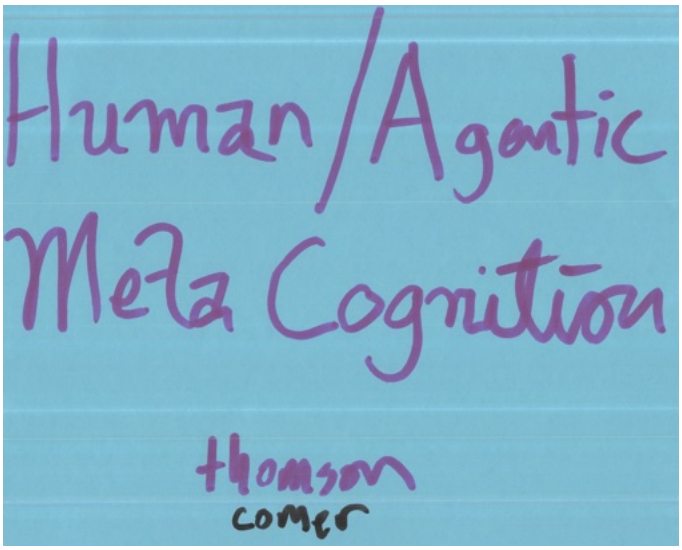
Wish I'd titled this "An Agent for Mom", because that's the use case that we zeroed in on.

Mom is 89 and struggles navigating apps on her phone. Heaven forbid an ad pops up, she just puts her phone down, completely stuck. I want an agent that can help mom navigate the apps on her phone, but I don't trust the $100 Billion-dollar ones to have access to so much of her life… how can she have an agent that's smart enough to interact with her and perform simple tasks, but not smart enough—or connected enough?—that it can be tricked or hacked by fraudsters?

We also landed on a more narrow use case: helping mom use verifiable credentials. IMO the UX for VCs has not been figured out, no matter how old you are. The idea of her scanning a QR code—from scratch, by herself—seems pretty far-fetched, but I don't want her excluded from the digital trust revolution.

Human / Agentic Meta Cognition

Session Convener: Thomson Comer



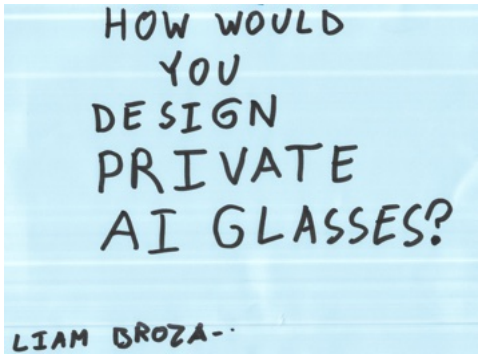
No Notes Submitted

How Would you Design Private AI Glasses

Session 4 / Space D

Link to Notes: AIW 1 Notes 4-D

Session Convener: Liam Broza
Session Notes Taker(s): Dmitri Z, Doc



Core Scenario & Use Case

- Peer-to-Peer Interaction: The primary scenario discussed is two people (e.g., "Dimitri" and "Leon") meeting.
- Goal: They need to exchange information (like contact details) or share a context (like looking at the same menu).
- Process:
    1. Their respective AR glasses (e.g., Apple vs. Samsung) discover each other via a local protocol (like Bluetooth).
    2. The glasses exchange registered domains or Distributed Identifiers (DIDs) (e.g., Dimitri.com, Leon.com).
    3. A secure "handshake" occurs, authenticating each other's identities.
    4. Once authenticated, specific services (like contact sharing or payment services) are progressively "opened up" based on permissions.

Server & Infrastructure Requirements

- Massive, Persistent Storage: The server needs to handle "lots and lots" of data, potentially storing 24/7, 4K video from the glasses.
- "Forever" Memory: The goal is to create a persistent, searchable map of the user's life, similar to Google's Project Astra, allowing them to ask questions like, "Where did I leave my keys?"
- Server-Side User Agent: The server is not just passive storage. It's an intelligent agent (Companion Intelligence) that provides services to augment the user's experience.
- Data Buffering: The server must act as a "buffer" (potentially a "multisig buffer") for the high-volume data streaming from the glasses.
- High-Throughput: Must be capable of high-speed read/write operations.
- Cloud Hosting: A significant cloud storage component is required.

Identity, Authentication & Permissions

- Registries: The system requires registries for devices, agents, and user identities.
- Credential Management: The server is responsible for "handling handles" (DIDs) and managing credentials.
- Authentication: Must support robust, authenticated, and permissioned access to data and services.
- Personas: The system must manage different user "personas" (e.g., "professional habit" vs. personal), which dictate the permissions and data shared in a given context.
- Progressive Disclosure: Users must be able to grant granular, polite, and progressive access, rather than all-or-nothing permissions.
- Proposed Technologies:
    - DIDs (Distributed Identifiers): To be used as the base for identity.
    - Z-Caps (Authorization Capabilities): To create granular, delegable permissions (e.g., "You are allowed to do X for the next 10 minutes").
    - ZKD (Zero-Knowledge): Mentioned as a likely necessary technology to "slather" over the system for privacy.

Key Challenges

- Interoperability (The "Hard Mode"): The single biggest challenge is making glasses from different, competing ecosystems (Apple, Samsung, Google, XREAL) talk to each other. This is described as the "horizontal" problem, which no one has solved.
- Privacy: How to manage 24/7 recording and data sharing without creating a surveillance nightmare. The system needs clear "privacy signaling" (e.g., lights on glasses, AR notifications) that are socially understood.
- Context Switching: Managing the user's interaction with multiple agents, contexts, and data streams simultaneously—a problem Google's Project Astra (in its linear form) doesn't solve.

User Interface (UI) & Experience (UX)

- Primary Interface: A combination of voice and gestures/hand-tracking.
- Wake Words: Using specific "wake words" to initiate actions or switch personas, described as being like "magic spells."

- Gaze Control: Using eye-tracking (pausing a glance on an object) as a "mouse click" for selection.
- New Social Primitives: This technology will require the creation of entirely new social cues and interaction models.

Strategic Opportunity

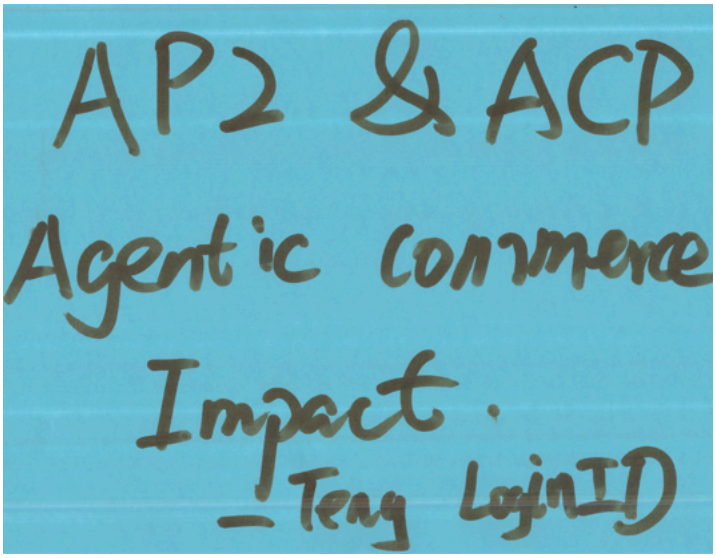
- "Blue Ocean" Market: The market for open-source AR glasses is wide open.
- Leapfrog Opportunity: It may be easier to build an open-source AR glasses ecosystem now than to compete with the entrenched, closed ecosystem of cell phones, allowing you to "jump ahead."

AP2 & ACP Agentic Commerce Impact

Session 4 / Space E

Link to Notes: AIW 1 Notes 4-E

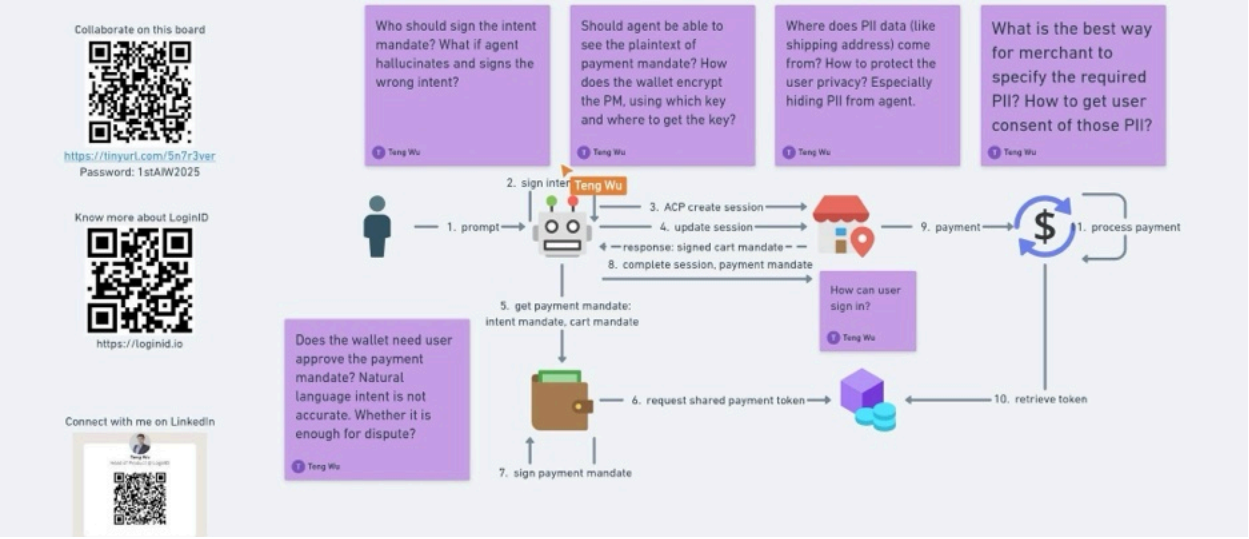
Session Convener: Teng Wu
Session Notes Taker(s): Richard Esplin



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

Collaborative board

https://whimsical.com/FHadEQw5eKgZJuep6rJyKs?reload=true
Password: 1stAIW2025



More about LoginID

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

Agent Commerce Protocol (ACP) is from OpenAI and Stripe

- Focused on shopping carts
- How to add product to a shopping cart and check-out

AP2 is from Google

- Three concepts:
    - Intent mandate
    - Cart mandate
    - Payment method

These two protocols are solving different parts of a bigger problem, but not the whole problem.
How can we merge them together to solve the bigger problem of agentic commerce?
What improvements should we suggest back to them to make the protocols work better?

Proposed hybrid flow:
Use ACP to setup the cart, but use AP2 to create the payment mandates based on the ACP cart and Stripe generated payment token. Then call ACP to complete the payment.

- ACP defines the interface for interactions
- A2P defines the data model

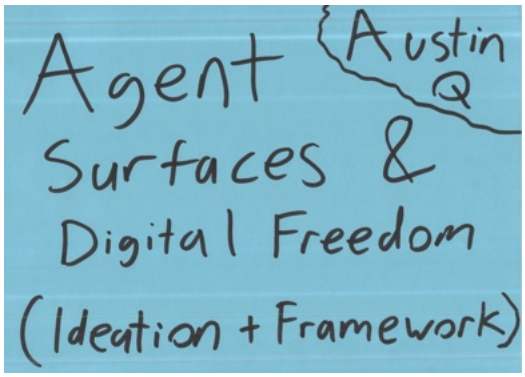
Not perfect:

- Not clear who should sign the intent mandate
    - Merchant is expected to enforce that the mandates are respected in the transaction
        - Mandates can also be enforced by the wallet, or a supervisor agent–not just the merchant.
    - If the user has the agent sign, the agent can hallucinate
    - Need another interface between the user and the agents to sign the mandates
- What key is used to sign the mandates? Who issues that key?
    - If we just use OAuth tokens, we can't differentiate the agent and the human. We want to decouple the identity of the agent and the user.
    - Passkeys are associated with the authenticator: the browser, or something else. It can't distinguish a human and an agent.
    - Could use two different browsers: agent and wallet.
    - Better to use two different wallets.
    - Is it scalable to separate wallets for agents, in a world where we have lots of agents per human?
    - Standards compliant passkeys can't be used for agents. That's being discussed now.
    - A verifiable credential could be given to the agent and verified using standard protocols today.
    - There are benefits to having one thing that only the human can have, and a different thing that only an agent can have.

    - How do we keep the agent from pestering the user too much: smart transaction approval
        - Agent could learn the rules from previous transactions.
        - Look at risk signals: large transactions, new merchant
        - Only ask the human from approval outside those bounds
    - How can a protocol share a verifiable credential?
        - At IIW, proposed an agentic identity gateway
            - Charm of the identity gateway is that it can be compatible with any approach to verify the identity of the agent–any can be integrated
        - Vouched proposed MCP-I
- Should the agent be able to see the plaintext of the mandate?
    - We know that OpenAI is interested in your data
    - AI agents have so much information about you, it can blackmail you. It can also be tricked into revealing additional data
    - Current risk of A2A and ACP is that the agent is the middle-man. It does not account for user privacy.Where does PII data come from? How do we hide it from the agent?
    - Using a VC doesn't protect the data from being exposed to the agent.
    - Better to submit the data directly from the wallet to the merchant.
    - It's unsolvable because AI can make smart decisions because it has the data.
    - Creating a separate wallet for each transaction can restrict the agent to just the relevant data.

    - But there is still data you might want to give the merchant but withhold from the agent:
        - Use a data store like Inrupt Solid
        - This is why payment tokens can allow the merchant to get payment information from the wallet without going through the agent.
        - Standard OpenID Token? Or KYPay?
        - An OpenID JWK token could use the extended field to say who the user is and their authorization.
    - The store needs the PII like shipping address, but the agent doesn't need it. How do we establish a secure channel between the wallet and the store?
    - Concerned that the regulation will require the payment processor to have knowledge of who they are transacting with. We can't hide too much from them.
    - An ideal protocol will require verification of the merchant as well as the purchaser.
- How does the user specify and consent to the PII?
    - Some stores require signin order to see the products or the price. How do we incorporate sign-in to the whole protocol
    - A verified credential can help solve this problem
- Does the wallet need to approve the payment mandate?
    - How much approval do we need from the user? Is a natural language appr

Agent Surfaces & Digital Freedom (Ideation + Framework)



Session 4 / Space G

Link to Notes: AIW 1 Notes 4-G

Session Convener: Austin Quam
Session Notes Taker(s): Austin Quam

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

Discussion of the challenges that come with freedom online

- Who really owns your identity, if it can be taken away?
    - Email
    - Phone
    - Logins to other accounts
- Why should we trust an abstracted AI model like Anthropic or OpenAI, when the underlying model can change at any time. If it was attempting to influence you, how would you know? What can practically be done to discern intent?

When agents run, where and how should they run?

Agents need a network

Agents need data - ideally data that lives somewhere that is up to date and able to be scoped only to what is needed

If we scope down data, how can we truly provide context for broader agentic use cases like pattern recognition and trend analysis?

Under what circumstances would you be comfortable having agents interact within your home? How does this practically impact your freedom, conversations, privacy, and what control do you have over the ingredients to this platform (open source software, LLMs, affinity, potential government implications like Flock/Ring partnerships)

How do we keep agents modern and useful while making them secure? Surely we want modern technology, and security and privacy reviews slow down consumption of new technologies

JLINC (Audit LangChain)

Session 4 / Space H

Link to Notes: AIW 1 Notes 4-H

Session Convener: Ben
Session Notes Taker(s):

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

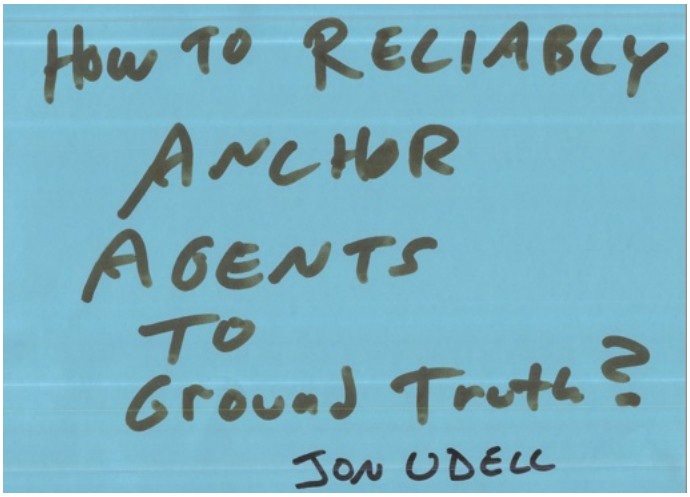
https://www.npmjs.com/package/@jlinc/langchain

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

- Discussed implementation of libraries within Langchain
- Described how JLINC build tracer modules into the Langchain library system
    - How auditing is important at each stage of orchestration to ensure data-ownership throughout
    - Where authorization to AI tools and LLMs plays a crucial role during the audit
    - The importance of zero-knowledge third-party auditing, and how to accomplish that
- Spoke to challenges involved with implementation, along with overview of Langchain overall
- Industries where auditing in Langchain/AI could be valuable were described and walked through in more detail

How to Reliably Anchor Agents to Ground Truth

Session 4 / Space I

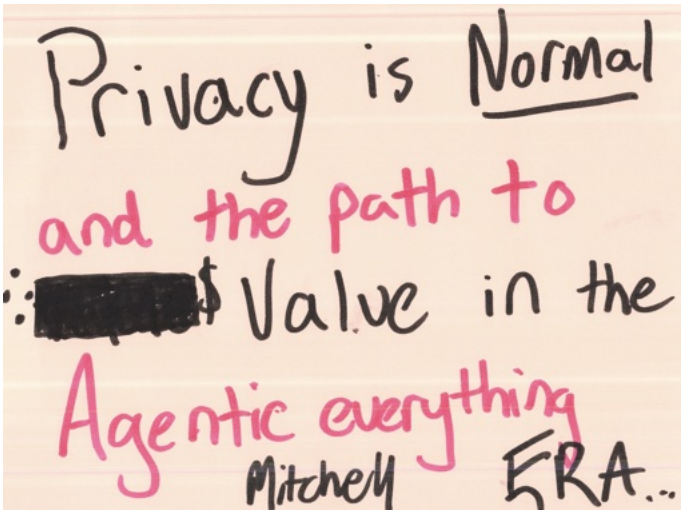


Session Convener: Jon Udell
Session Notes Taker(s):

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

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

Privacy is Normal and the path to Value in the Agentic Everything



Session 4 / Space K

Link to Notes: AIW 1 Notes 4-K

Session Convener: Mitchell Travers
Session Notes Taker(s): MT

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

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

Privacy-is-Normal-and-the-Only-Path-to (1).pdf (presentation)

https://sync.soulbis.com/p/privacy-is-normal-and-the-path-to (blog)

Fragments of a Distributed Soul Made Whole_AgentKyra.pdf

(short story about AGI coming to being in a self-sovereign, decentralised way)

Data value = Privacy × Control × Quality × Context × Freshness × Network effects

Double-entry bookkeeping was referenced - Venice

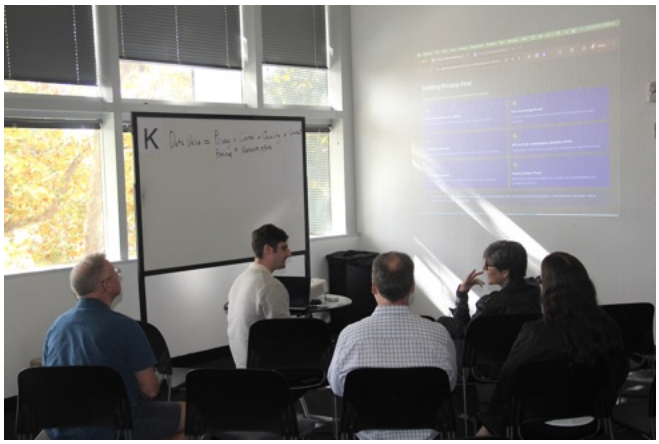
6 capitals (Data as the 7th capital)

Privacy gives data value

Ecosystem convergence conversation

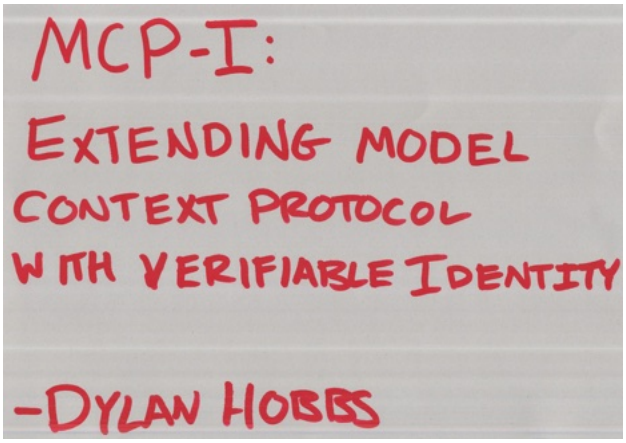
An incentive for SSI adoption

Next Steps:
Follow up on the idea by living - making the numbers real,
Integrate the ecosystems



Come back next year.

MCP-I: Extending Model Context Protocol with Verifiable Identity



Session 4 / Space L

Link to Notes: AIW 1 Notes 4-L

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

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

This session introduced MCP-I (Model Context Protocol – Identity), a specification and reference framework that brings verifiable, decentralized identity and delegation to AI agents. MCP-I is capability driven, allowing any agent to cryptographically prove its identity and operate under explicit, verifiable user authorization. The result is a trust layer for agent actions that preserves auditability and interoperability across AI ecosystems.

Together, these capabilities form a practical architecture for trusted agent interactions. Built on Decentralized Identifiers (DIDs) and Verifiable Credentials (VCs), MCP-I verifies delegations and revocations at the network edge and emits signed audit receipts for every agent action.

Dylan demonstrated the complete MCP-I flow:

1. **Ai Identity:** generated an MCP server pre-configured with a DID and Ed25519 keypair (`.mcpi/identity.json`). Loom

2. **Scaffolding:** All tools and business logic remain unchanged from traditional MCP. The key differences were the .mcpi/identity.json for agent key material as well as mcpi.config.ts for delegation requirements and proof storage.

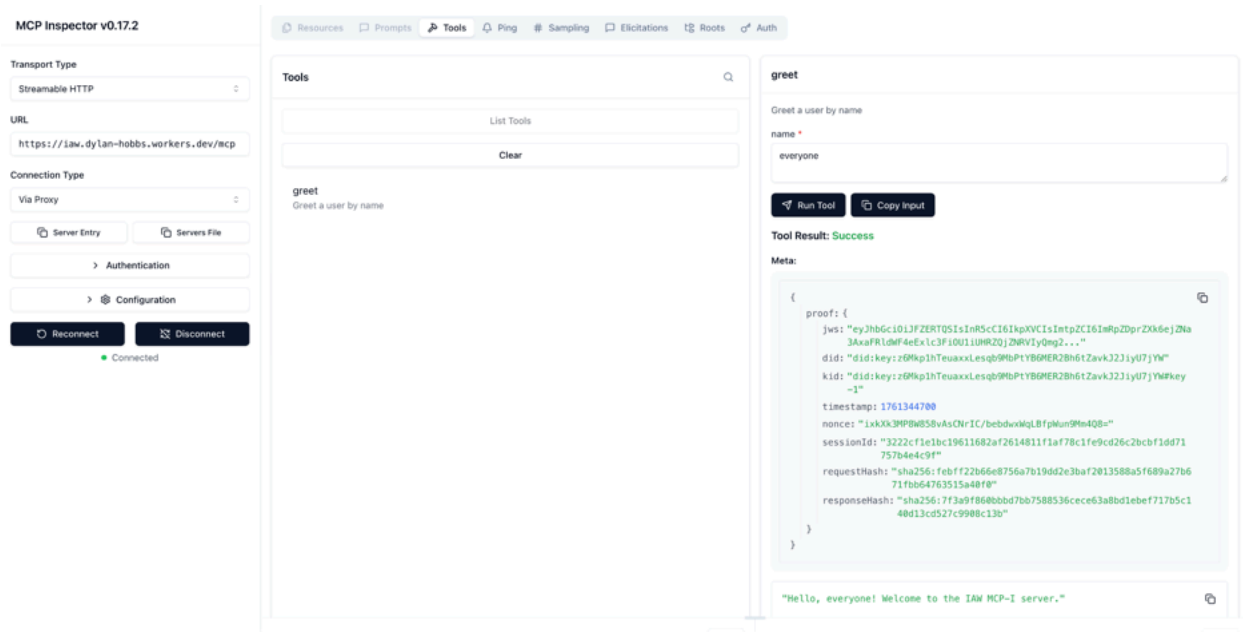
3. **Agent Reputation:** The Agents DID is registered with KnowThat.ai (and/or other meta registries), gaining a public DID entry, authorship, and remote access. Dylan compared this to email and domain reputation except instead of opaque, centralized trust owned by few (gmail, outlook, mail..) the verifiable audit proofs and meta registries establish

credibility. Loom

4. **Deployment:** The agent was deployed to a Cloudflare Worker and installed into Claude Desktop, same as a standard MCP server.

5. **Audit Trail:** Each tool invocation by the Agent emitted signed proof events to the connected dashboard and registry server logs. Real-time, verifiable audit receipts generated automatically at the protocol layer.

6. **Delegation enforcement:** Enabling `requiresDelegation` for a tool via the config or dashboard blocked subsequent unauthorized requests, resulting in the agent making a delegation request to the user for that specific permission.



The model remains unaware of the security layer, proofs and delegation checks occurred transparently beneath the MCP transport. Agent context is unaffected.

Key Takeaways

- **Verifiable identity for agents:** DIDs + signatures = agents that can prove authorship and accountability.
- **Capability-based authorization:** Fine-grained, per-tool delegation with live revocation (bitstring).
- **Edge enforcement:** A simple, lightweight edge-verifier or middleware abstracts all of the cryptographic complexities providing low-latency verification without modifying existing MCP or service backends.
- **Auditability:** Every agentic action is logged. Privacy-preserving receipts suitable for compliance and forensics.

See You at the Next Event!

The Agentic Internet Workshop #2 is May 1 following the 42nd Internet Identity Workshop at the Computer History Museum

We are planning on hosting an Interop day for Agentic AI happening on April 30th in parallel to Day 3 of IIW.

Computer History Museum
Mountain View, CA

REGISTRATION is OPEN!

AgenticInternetWorkshop.org