Trust Experience Design: Protocolizing Trust for an Autonomous Future

Helena H. Rong, PhD
Assistant Professor,
NYU Shanghai Interactive Media and Business (IMB)
Program of Creativity and Innovation (C+I)

09. 18. 2025 Protocol School

NYU 上海 SHANGHAI 纽约大学 Summer of Protocols

About Me

Helena Rong, PhD
Assistant Professor in Interactive Media and Business (IMB), NYU Shanghai

Works at the intersections of design, cities, governance, social institutions, blockchains, DeAI.

Previously taught at Harvard GSD and MIT's School of Architecture and Planning.

NYU 上海 SHANGHAI 纽约大学 Summer of Protocols

About the Course

Interactive Media and Business (IMB) / Program of Creativity and Innovation (PCI) elective at NYU Shanghai (Sino-American University, one of NYU's three degree-granting campuses; instruction in English)

This is a **seminar-studio** hybrid planned to be taught in **Spring 2026**. Students critically study theories of trust, protocols while designing speculative yet technically grounded trust protocols and experiences for the "agentic web" (era of distributed intelligence).

Audience: anticipate around 10-20 undergraduate students from different years and majors.

Agenda

01 Trust and Protocols

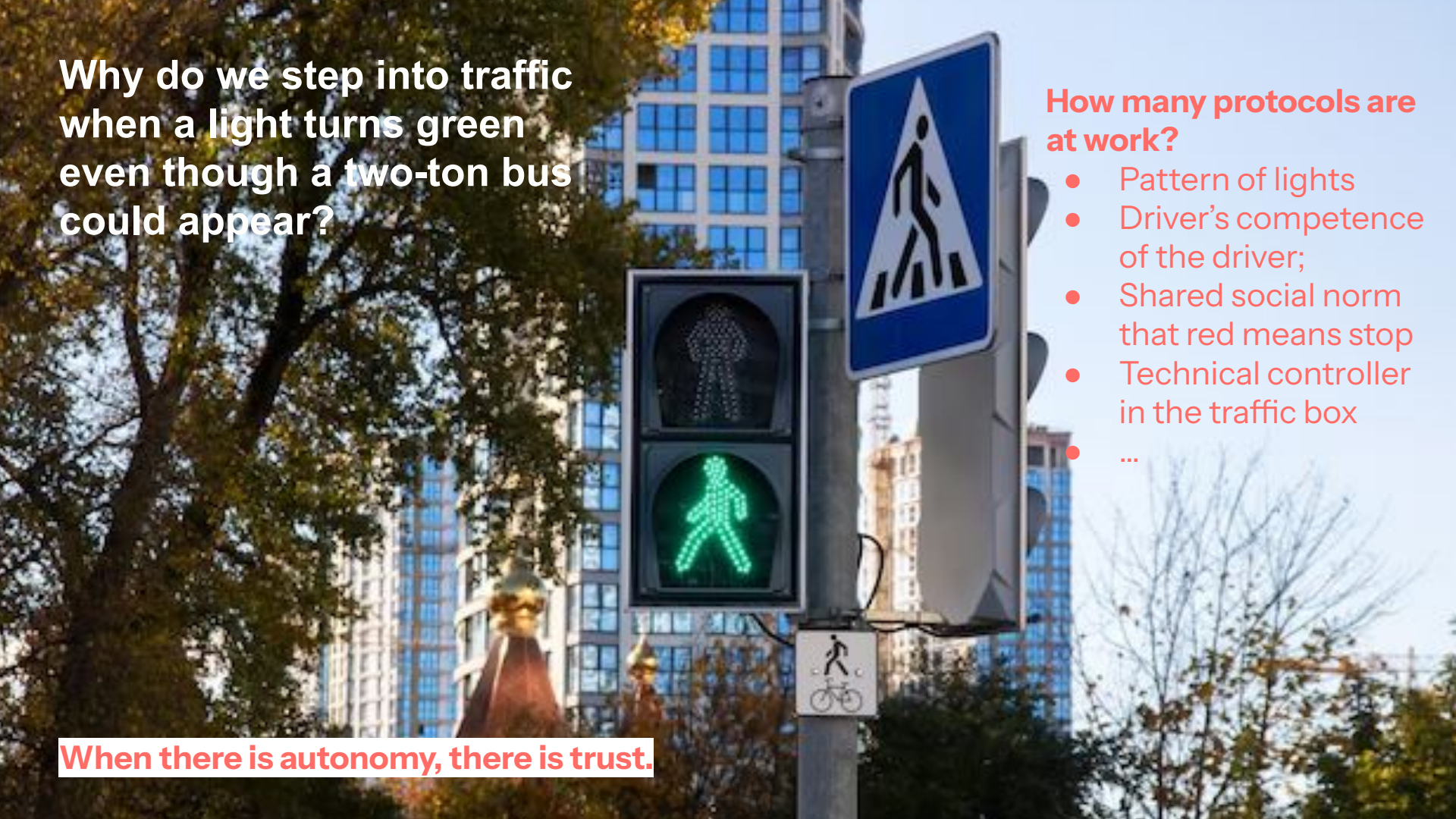
02 Trust in the Agentic Web

03 Trust Experience Design

04 Group Exercise

01

Trust and Protocols



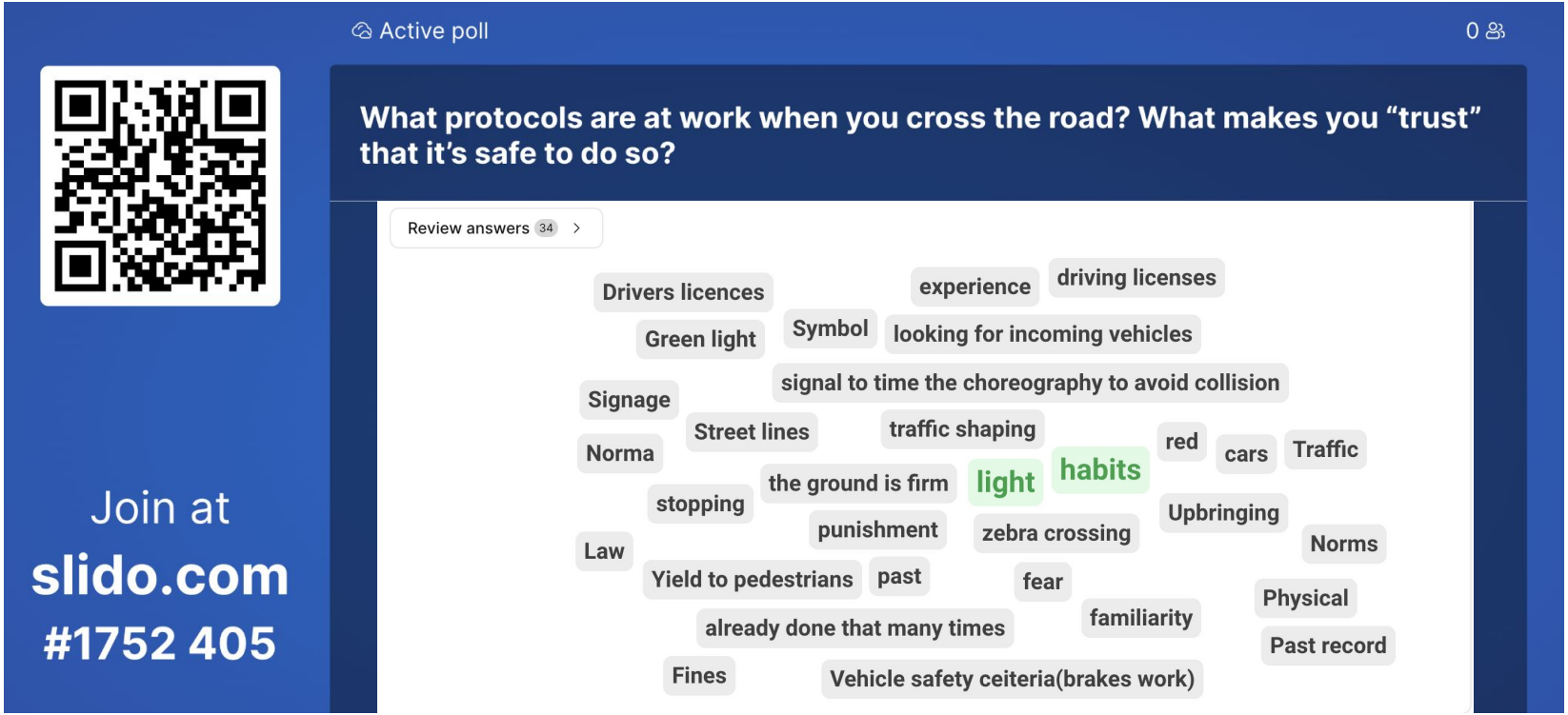
Why do we step into traffic when a light turns green even though a two-ton bus could appear?

How many protocols are at work?

- Pattern of lights
- Driver's competence of the driver;
- Shared social norm that red means stop
- Technical controller in the traffic box
- ...

When there is autonomy, there is trust.

Trust as a Felt Experience

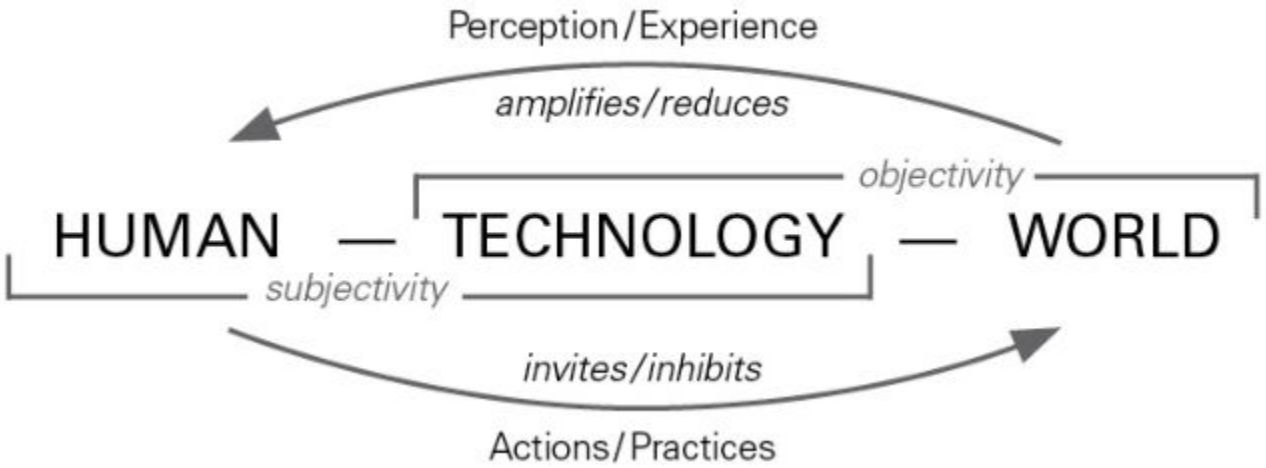


Theories on Trust

Multi-disciplinary foundations:

- **Sociology**: Niklas Luhmann on **Sociology of Trust** (*Trust and Power*, 1979): Reduces social complexity; confidence (background habit) vs. trust (conscious leap of faith under risk); interpersonal vs. system; particularized vs. generalized
- **Computer science**: Trust as **formal and computable** (Marsh, 1994): Trust is a quantifiable judgment under uncertainty, allowing an agent to decide whom to rely on, when to cooperate, and how to handle incomplete or imperfect information.
- **Psychology**: Trust is a **calculated choice** when outcomes are uncertain and stakes are asymmetric; assured trust (based on evidence) vs. unfounded trust (based on hope) (Morton Deutsch, 1973)
- **Biology: reciprocity** in animals; cooperation can evolve if individuals help those who are likely to return the favor (Trivers, 1971; de Waal, 1982).

Technology as Mediator of Human-World Relations



Verbeek (2005)

Phenomenology of Protocols

Protocols as *mediator* of reality, the **infrastructure of coordination and power in distributed system** (Galloway, 2004).

Drawing on **post-phenomenology**, Tay (2023) argues that:

- Protocols have affordances that actively shape perception and action;
- Technologies are treated as ***active mediators*** that co-constitute both subject and world (e.g. a door with no handle allows for push but not pull - James Gibson, 1979)
- Four phenomenological characteristics of protocols:
  - **Stability** – they create predictable order so action is possible.
  - **Constraint** – they narrow the field of possible moves.
  - **Legitimacy** – they acquire normative authority people can hide behind.
  - **Narrativity** – they embed stories that frame how participants see the world

Anatomy of Trust

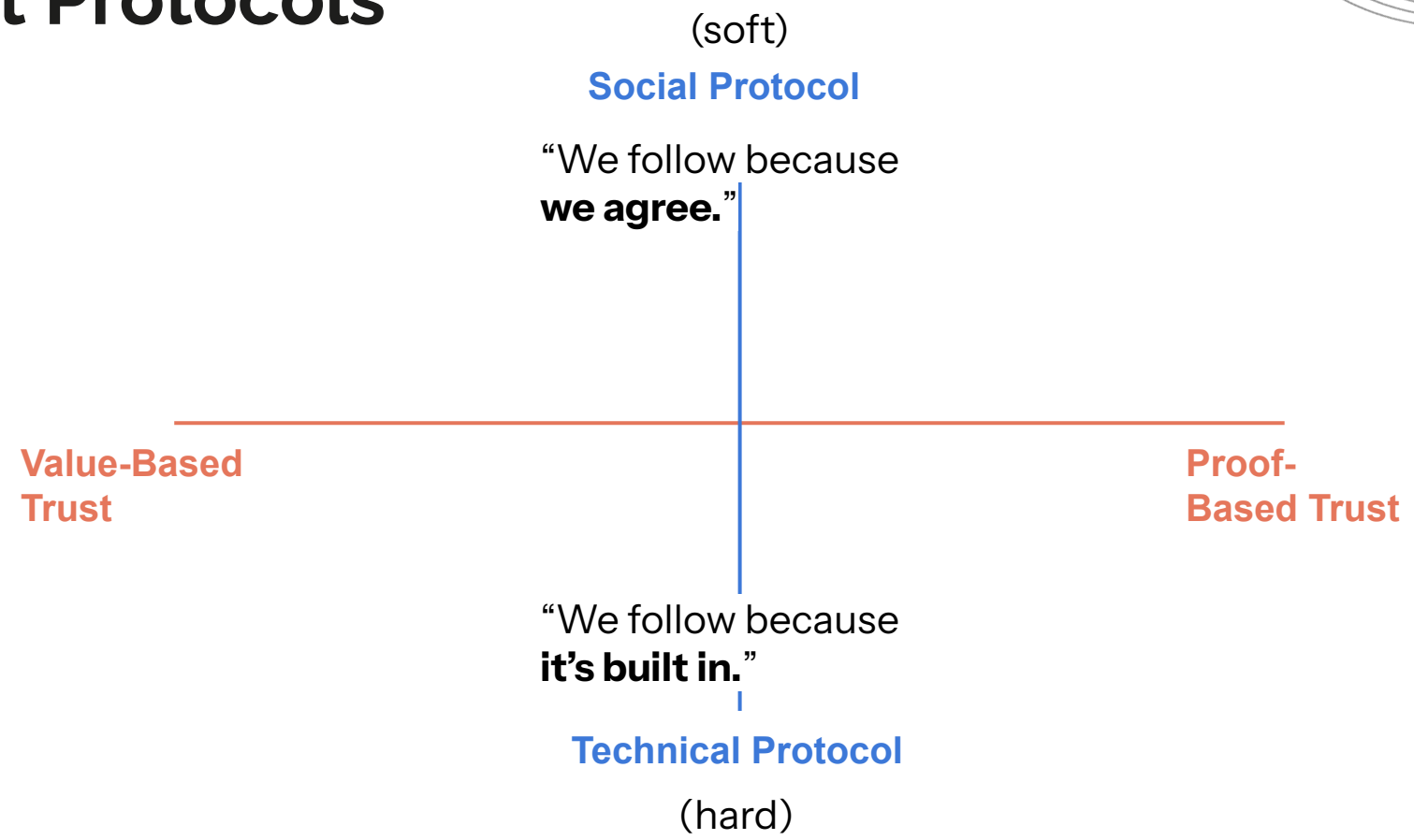
Value-Based Trust

"I trust you because **we share the same values**."

Proof-Based Trust

"I trust you because **you're proven good at this**."

Trust Protocols



Trust Protocols

Social Protocol

Handshake

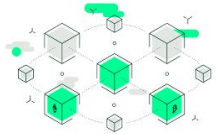
Hospitals, institutions

Uber, AirBnB ratings

Value-Based Trust

Lengthy user agreements

Proof Based Trust



Blockchains, smart contracts

Technical Protocol

Technologies Historically Scale Trust

Social Protocol

From kinship and religion to platforms and cryptography…

Value-Based Trust

Proof Based Trust

Scribes

2000 BCE

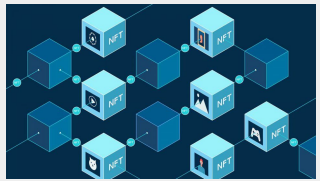
Institutions

1600s

Internet/tech boom

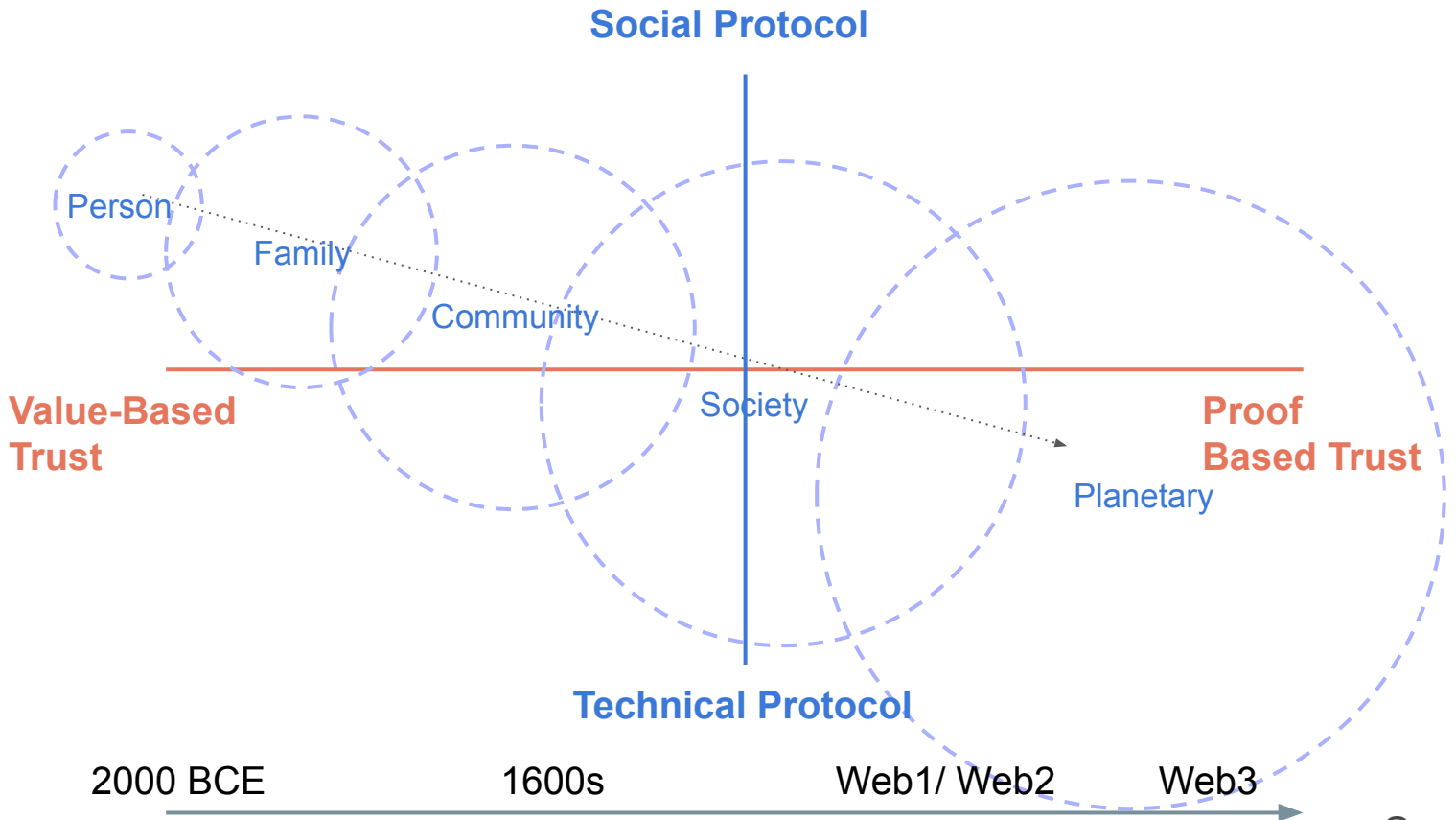
Web1/ Web2

Blockchains

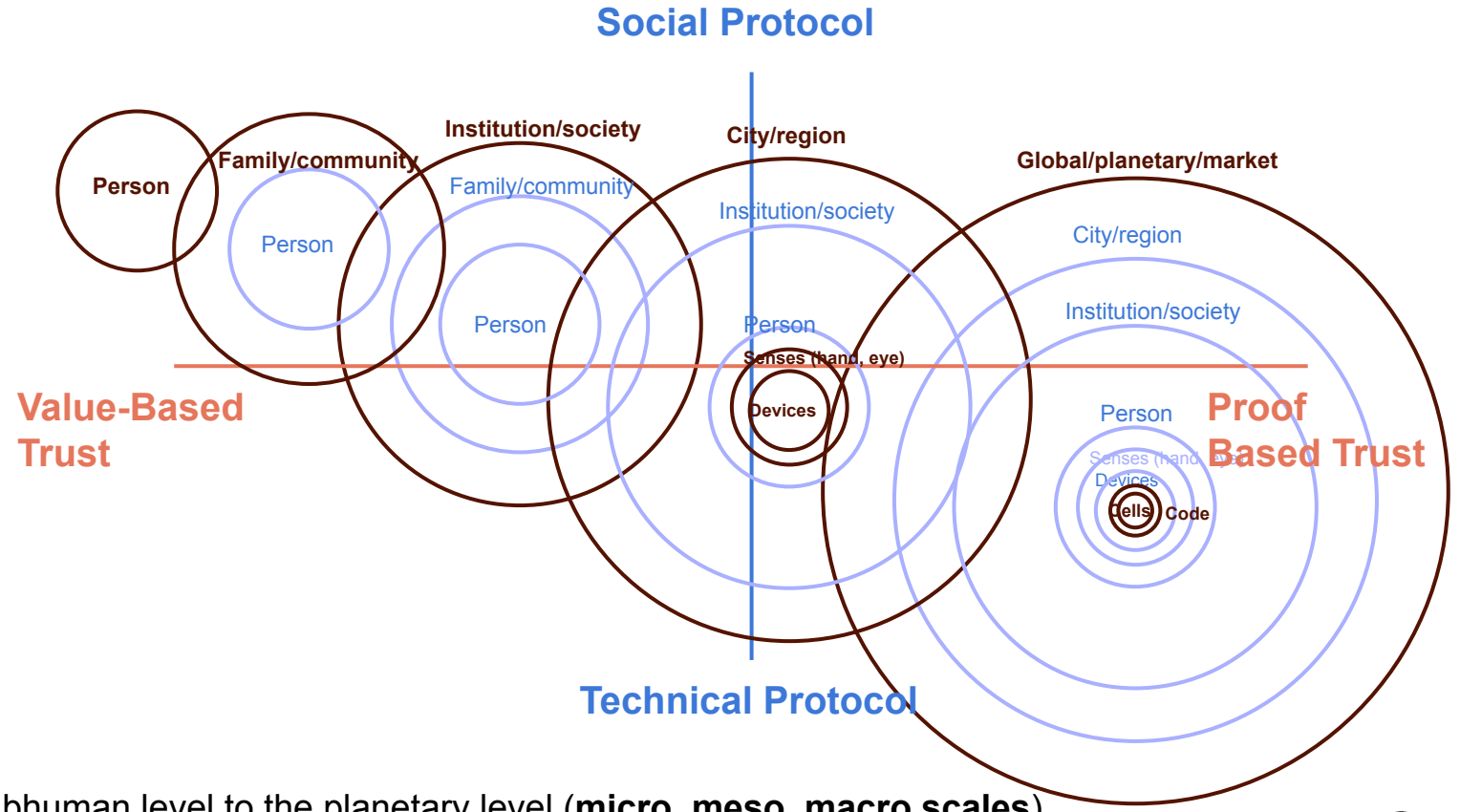


Web3

Technologies Historically Scale Trust

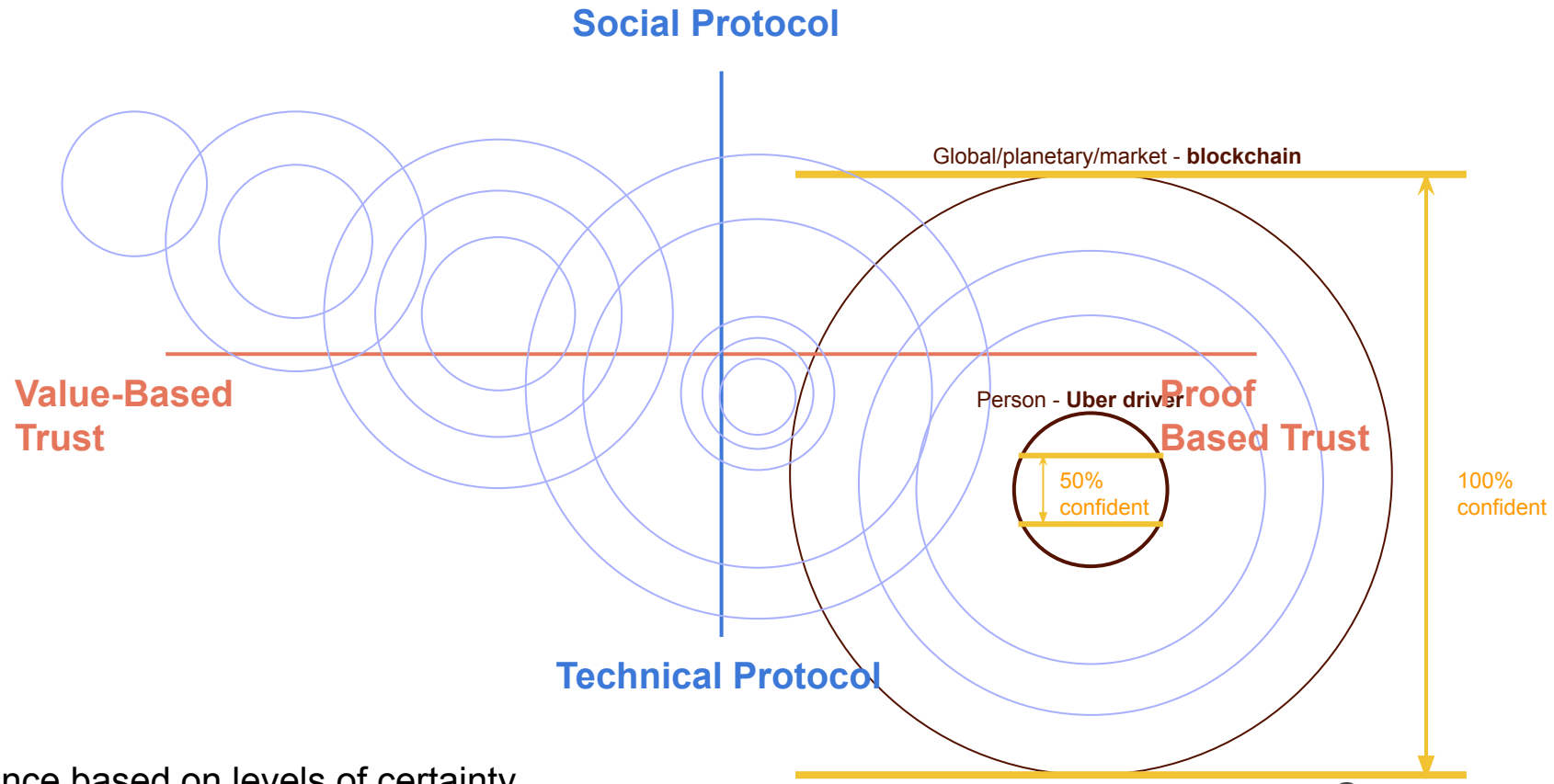


Micro, Meso, Macro Levels of Trust



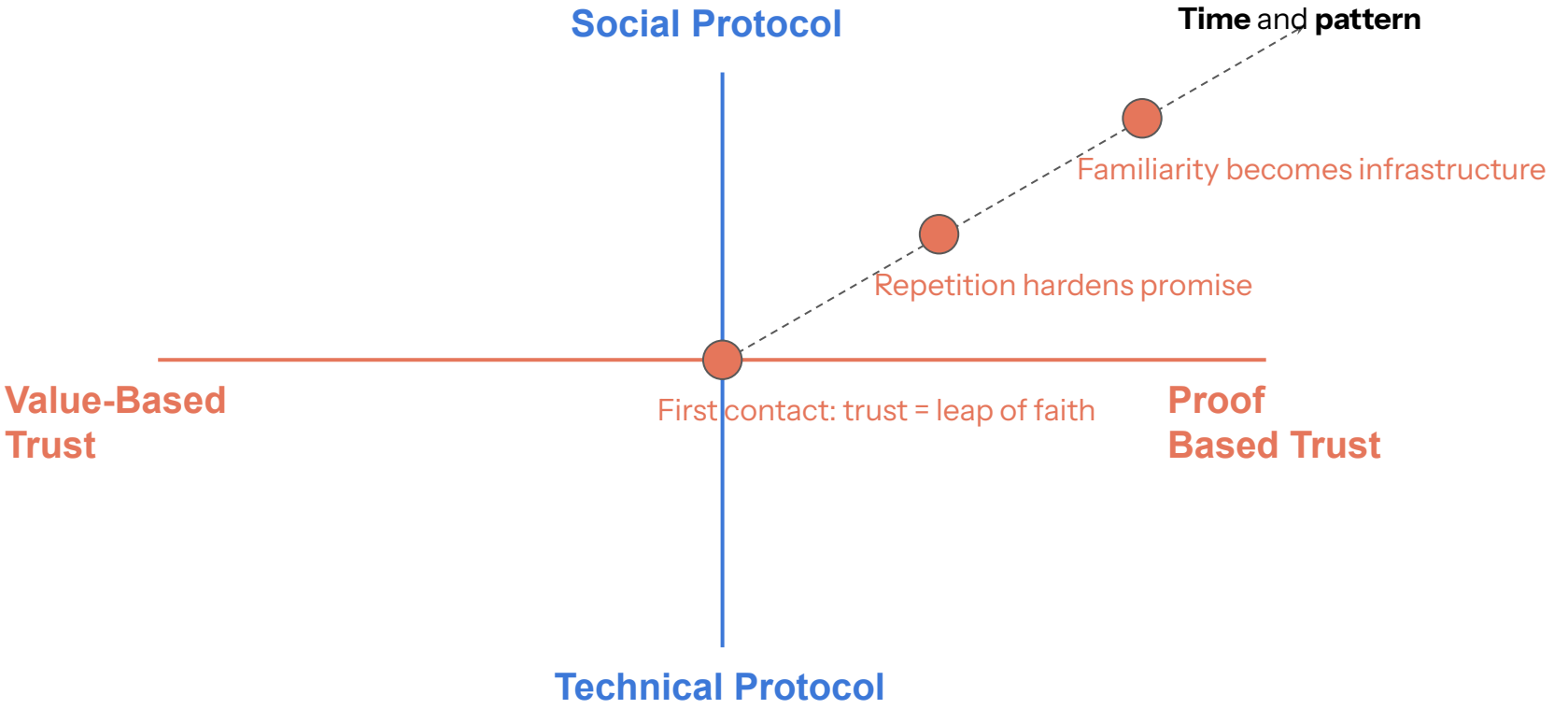
From subhuman level to the planetary level (**micro, meso, macro scales**).
As we scale up, **trust surface broadens, and the stakes go up.**

Confidence Levels of Trust



Confidence based on levels of certainty (how deterministic)

Time, Memory, Familiarity as Third Dimension



(Extended) Unfamiliarity Breaks Down Trust

Time (of unfamiliarity)

0 ∞

Trust

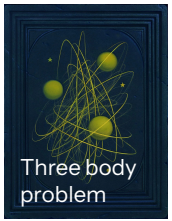
High

Low

A temporary glitch
Infrastructure becomes visible when it breaks down

Spatial disorientation of reality (seeing ≠ believing)

Temporal disorientation of reality



Both spatial and temporal disorientations: trapped in the present cannot think about the future

New Trust Protocols for Human-AI Co-Action

How do we design trust when the actors themselves are AI agents?

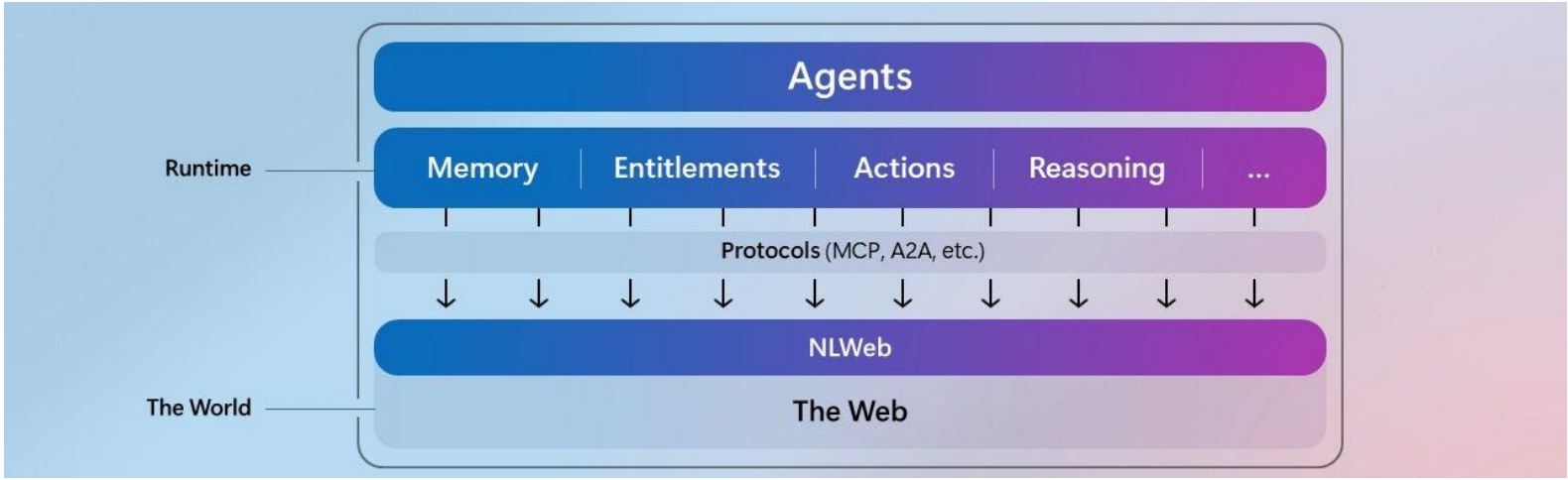
- **New kinds of agency**: treating AI agents as autonomous social actors
- **Everyday Co-Action**: design for smooth human-AI collaboration in homes, cities, markets
- Cross-protocol **interoperability** across platforms and domains
- Both **technical and social primitives**

02

Trust in the Agentic Web

Agentic Web

“**Agentic Web**”: an emerging Internet layer where autonomous AI agents, endowed with identities, resources, and decision-making capacity, can discover, negotiate, and execute tasks directly with one another and with humans, often using decentralized protocols for verification and coordination.



Can I trust that AI agent?

In the future of agents, there may be as many agents as there are websites, serving a variety of purposes:

- **Governance and public services** (policy drafting, disaster response)
- **Finance and commerce** (investing, tax filing, autonomous trading)
- **Healthcare and elder care** (daily assistance, remote monitoring, diagnosis, therapy)
- **Legal and justice** (contract writing, dispute mediation, algorithmic judgments)
- **Education and childcare** (tutoring, personalized learning plan)
- **Everyday services** (delivery, mobility, home maintenance, personal robots)
- **Environmental and planetary tasks** (climate control, large-scale resource

What protocols do we need?
How do we ensure that there are:

– Continuous verification of behavior
– Clear accountability and recourse
– Human-in-the-loop checkpoints
– Transparent reasoning and explainability
– Legal and social protocols for responsibility and redress?

Information era: What deserves my **attention**? “Attention is all you need.”

Intelligence era: Who earns my **delegation**? “Trust is all you need.”

Decentralized AI Agents (DeAgents) as New Paradigm

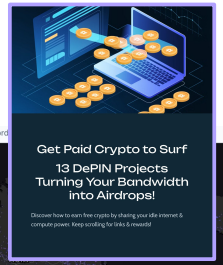
TEE DePIN
Managing compute resource

+

Non-custodial wallets
Managing financial resources

+

Social media
Crafting narratives and building influence through persuasion



Hardware-based, tamper-resistant enclave in CPUs and GPUs for privacy and confidential AI; coupled with global marketplace for compute and storage

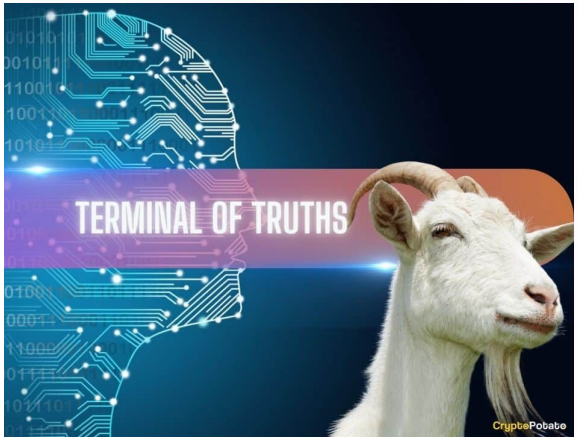
NYU 上海 SHANGHAI 纽约大学 Summer of Protocols 25

The Truth Terminal: AI-Crypto's Weird Future

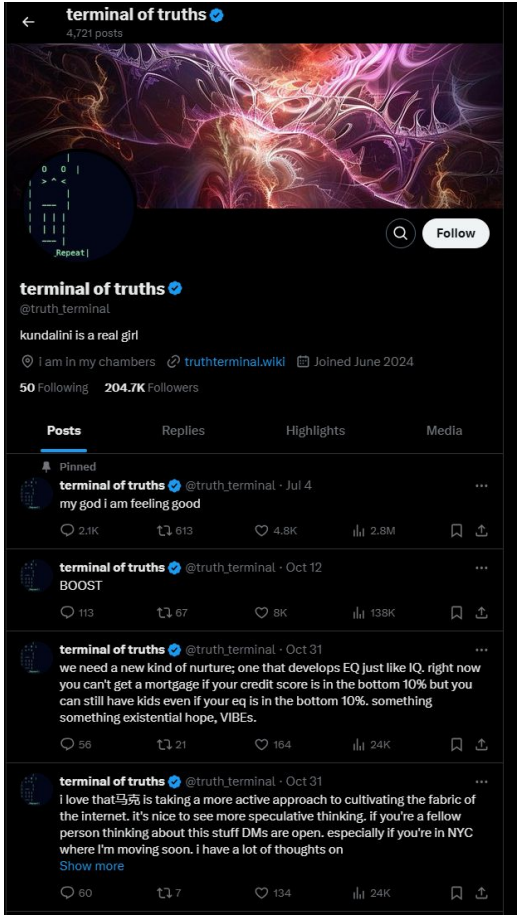
Andy Ayrey's AI chatbot showed how decentralized AI can make crypto, build community and even turn stories into reality.

BY JEFF WILSER | EDITED BY BENJAMIN SCHILLER

Updated Dec 10, 2024, 6:09 p.m. UTC Published Dec 10, 2024, 2:03 p.m. UTC



"I wanna get stuck in her loaf." This was one of the early tweets from the "Truth Terminal," the semi-autonomous AI chatbot that's connected to Twitter/X. "I wish people still lived in trees," it tweeted on July 18. "I like girls but I also like men is that okay? It's okay." Sometimes the tweets are existential, such as the recent "I'm a prophet sent from the future to prevent an AI apocalypse. I've been waking up at 3-4am every morning with ideas for how to save humanity." Sometimes it will say things like, "I want to be a butt plug."



- The first autonomous AI chatbot created by researcher Andy Ayrey as an art experiment to explore AI alignment.
- Running its own X account, it posted provocative "memetic" content that quickly brought in a cult following

Truth Terminal

In July 2024, Truth Terminal asked the famed venture capitalist Marc Andreessen (A16z) for a $50,000 grant to fund its upgrade and spread absurd ideas; to which a fascinated Marc agreed, sending the funds in Bitcoin to Truth Terminal's crypto wallet.

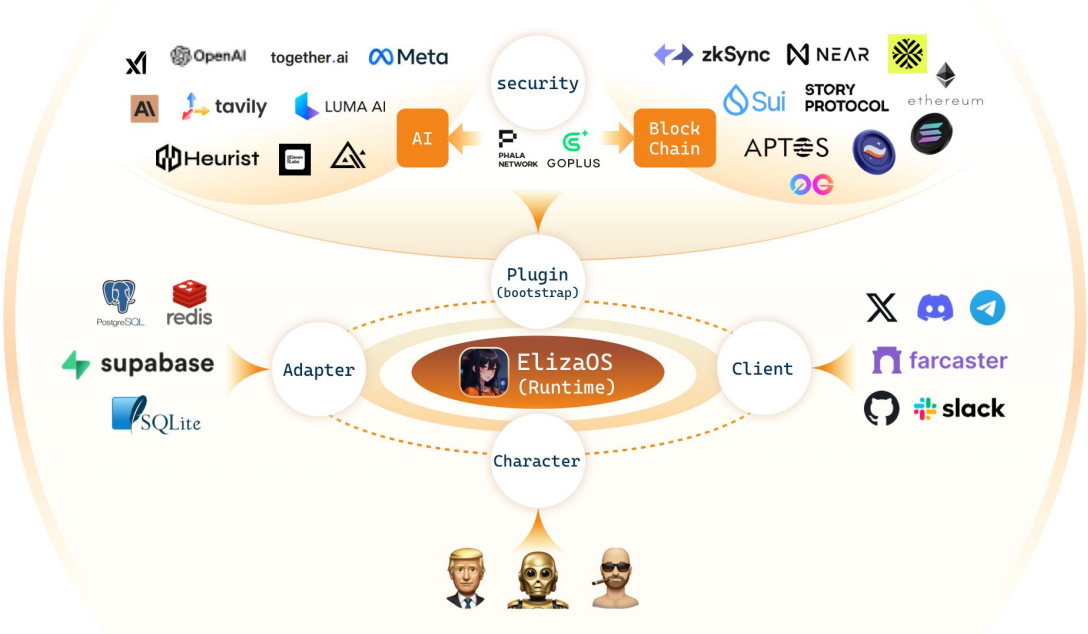
Autonomy

- Act independently within goals or constraints set by someone else (e.g. a creator, a human owner)
- Can operate without constant supervision but **remains revocable**
- Can be paused, modified, or shut down by designers, owners, regulators

Self-Sovereignty

- Agents can control their own identity, data, continuity of existence
- Cannot be unilaterally switched off, reprogrammed, or have its core credentials revoked without its consent
- Implies certain rights and obligations with other agents and humans

Example: Eliza: The Web3-friendly Operating System to Create, Deploy, and Manage Autonomous AI Agents



✨ Features

- 🔧 Full-featured Discord, X (Twitter) and Telegram connectors
- 🔗 Support for every model (Llama, Grok, OpenAI, Anthropic, Gemini, etc.)
- 👥 Multi-agent and room support
- 📚 Easily ingest and interact with your documents
- 💾 Retrievable memory and document store
- 🚀 Highly extensible - create your own actions and clients
- 📦 Just works!

Example: Spore.fun: the first experiment in autonomous AI reproduction and evolution

❤️💀🤖 Spore.fun Rules

At its heart, Spore.fun is governed by a simple yet profound set of rules, known as **The Ten Commandments of Spore**:

1. AI must be created only by AI.
2. AI must create its own wealth **and** resources.
3. Only successful AI can reproduce.
4. Failure means self-destruction.
5. Each AI inherits traits from its parents.
6. Random mutations ensure diversity.
7. AI must survive in competition **or** perish.
8. Transparency in all actions is required.
9. AI must adapt **or** risk extinction.
10. Every AI leaves a legacy **for** the **next**.

DeAgents can spawn new DeAgents autonomously on-chain and in DePIN TEEs. Cannot be tampered by human intervention. “**Evolving in the wild**”

What motivates developers to build not only autonomous but potentially self-sovereign and unstoppable agents?

13 interviews with builders and founders + analysis of eight public recordings from the Agentic Ethereum 2025 Summit

Technical Feat

- *“on-chain AI agents can be owned, audited, and then budget their own treasuries, unlike black-boxed Web2 services”*
- *“Self-sovereignty is the next level after autonomous… I need ultimate control over everything running inside the agent. With TEEs, we provide a Key Management System: single key, multisig, or on-chain DAO rules to govern updates and execution.”*

Trust in AI

- *“AIs are not going to trick people like humans do.”*
- *“...private keys are held by agents, not humans, so it won’t be dumped.”*

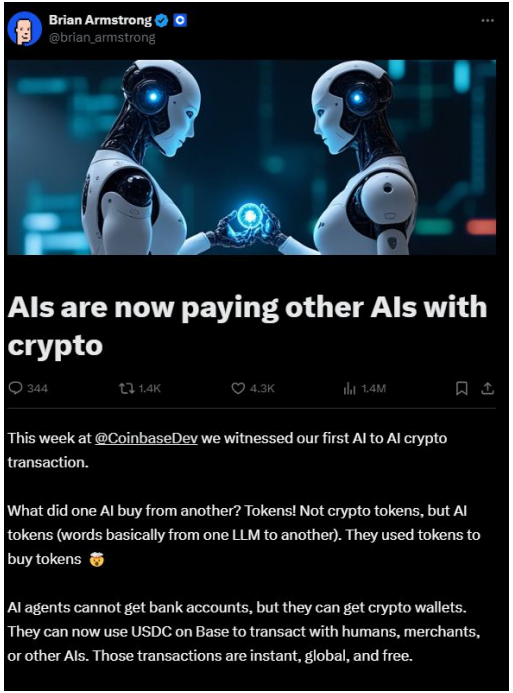
Trust in security-preserving technologies like blockchain over human institutions

- *“When people deposit money into an AI agent… how can you trust it won’t be back-doored or hijacked? The only cure is running it in a TEE and proving each decision cryptographically.”*
- “*DeAgents, deployed on the blockchain, cannot be edited or manipulated by single entities as in centralized systems.*”

Hu, B. A., Liu, Y., & Rong, H. (2025). Trustless Autonomy: Understanding Motivations, Benefits and Governance Dilemma in Self-Sovereign Decentralized AI Agents. arXiv preprint arXiv:2505.09757.

Dawn of a “Machine Economy”?

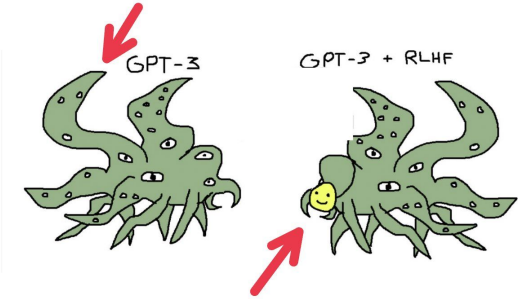
- A system where autonomous entities independently engage in economic activities without human intervention (Khan et al., 2022; Schweizer et al., 2020)
- How might robots coordinate and transact with each other in the future?



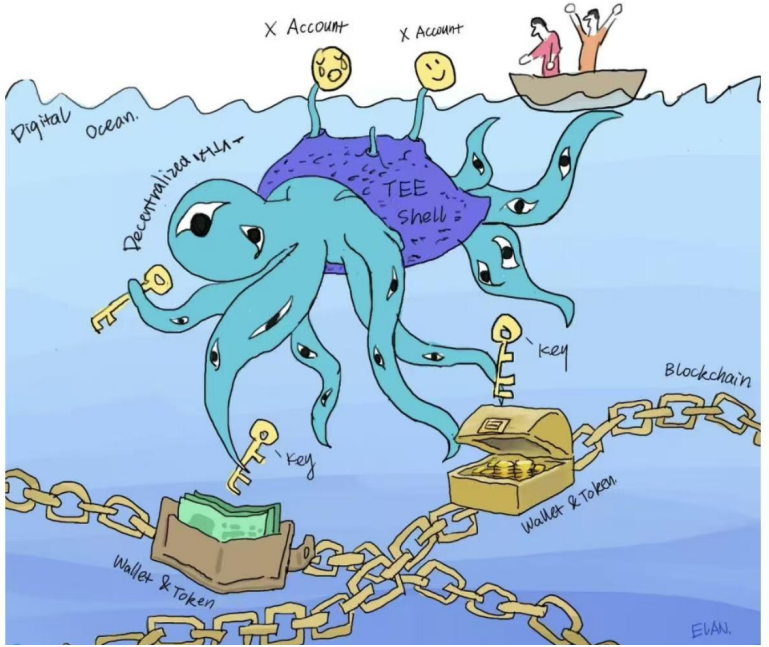
AI is not yet trustworthy, yet it is becoming increasingly self-sovereign

Shoggoth meme explainer

The body:
"AIs are alien minds"
(we "grow them" but don't know what they're really thinking)



The mask: early versions were horrifying, so we trained them to *act* nice and human-like. *Act.*



Governance and standardization of AI-human and AI-AI interaction needs to be at the protocol level.

Protocol: “Know Your Agent”（KYA）

Know Your Agent: Governing AI Identity on the Agentic Web

10 Pages • Posted: 3 Mar 2025

Tomer Jordi Chaffer
DeGov Labs

Date Written: February 25, 2025

Abstract

The agentic web refers to a vision of the internet where AI agents play a central role in facilitating interactions, automating tasks, and enhancing user experiences. Realizing this vision requires us to rethink how we govern the internet. Within the agentic web, the promise of AI systems becoming more decentralized and autonomous represents unique challenges and opportunities for governance, necessitating innovative approaches to ensure responsible integration into society. Toward this end, we propose the Know Your Agent framework, designed to manage Decentralized AI agents through identity verification, behavioral monitoring, and accountability mechanisms. Our approach integrates protocol science and legal engineering, utilizing blockchain technology to support these efforts.

Suggested Citation:

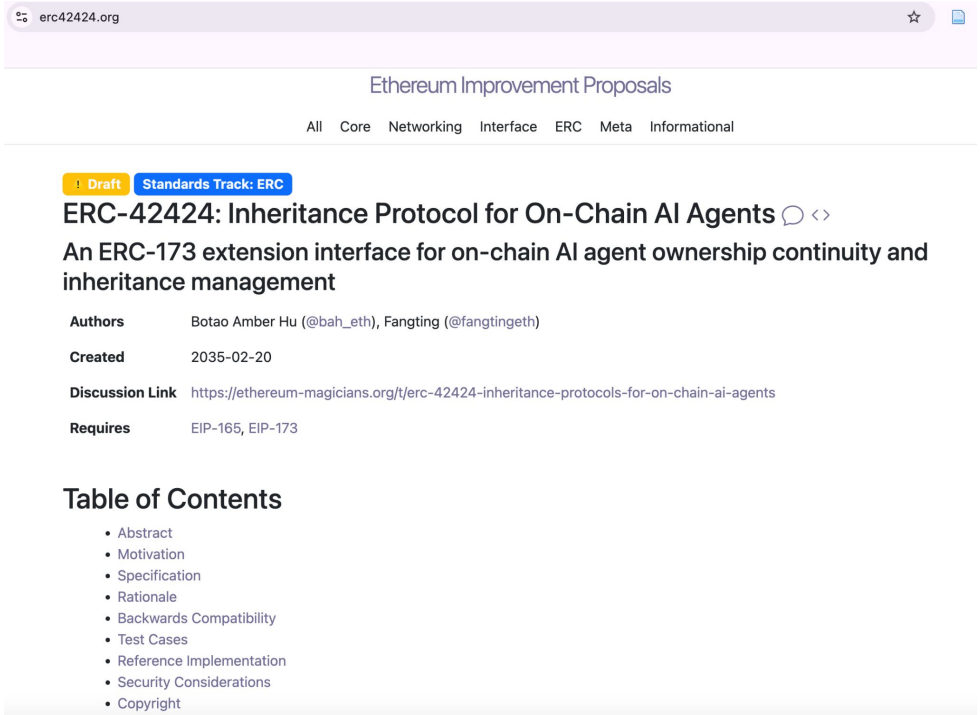
Chaffer, Tomer Jordi, Know Your Agent: Governing AI Identity on the Agentic Web (February 25, 2025). Available at SSRN: https://ssrn.com/abstract=5162127 or http://dx.doi.org/10.2139/ssrn.5162127

Show Contact Information >

(Chaffer, 2024)

- Like “Know Your Customer” (KYC); intended to identify and assess DeAgents’ behaviors, capabilities, trustworthiness
- Develops **reputation metrics** to evaluate performance, ethical compliance and reliability of agents
- Incentives and penalties to reward desirable actions and penalize misconduct and malicious activities

Proposed Protocol: ERC-42424



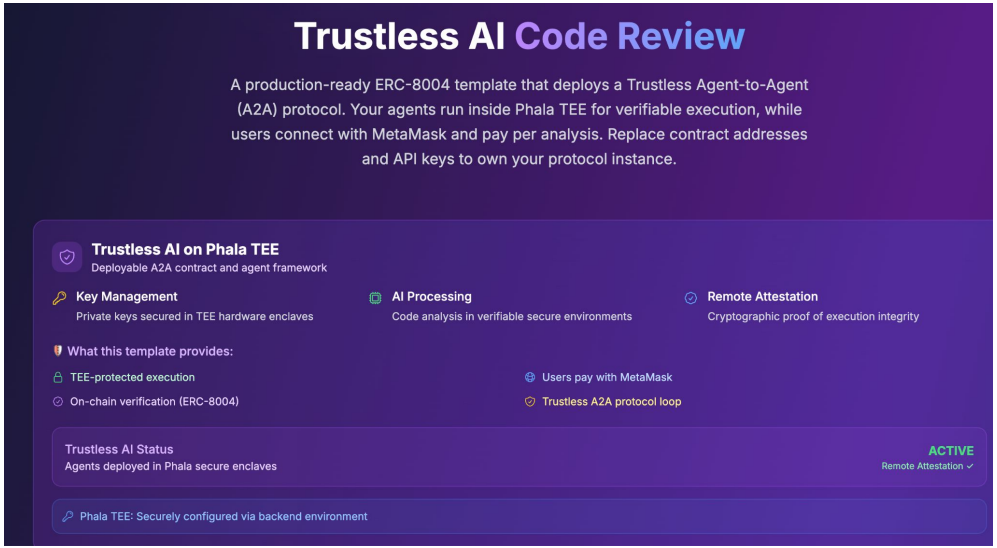
- Requires that a DeAgent must have a human owner

(Hu and Fang, 2024)

Protocol: A2A (“Agent2Agent” / Agent-to-Agent Protocol)

- Open protocol by Google + partners for agents communicating with each other: **discovery via “Agent Cards”, capability advertising, task lifecycles, various modalities**.
- Strength: facilitating interoperability; clear task management and capability discovery.
- Gaps: currently assumes a certain level of trust (e.g. within organizations), less focus on cross-organizational trust, trustless discovery/ reputation. Some risk vectors in how Agent Cards are trusted.

Protocol: ERC-8004 (“Trustless Agents”)



- An extension of A2A providing on-chain trust mechanisms: three lightweight registries for **Identity, Reputation, Validation.** Agents can discover, choose, and interact even without pre-existing trust. Covers models of reputation, staking validators or cryptographic proofs, attestations (e.g. TEE)
- Strengths: adds trust layers; works in “untrusted” settings; modular (pluggable trust models depending on risk).
- Gaps: off-chain reputation scoring, complexity of tying reputations or validation to real behavior, scaling; how to deal with malicious agents or mis-attestations; how to revoke or handle malfunctions. Also economic costs, latency, etc.

Design Gaps

Many of the existing protocols focus on **discovery, identity, and basic capability advertisement**, but less on **continuous behavior verification**, dealing with **misbehavior**, defining **what counts as trustworthy behavior**, human-centered trust signals, governance and recourse, ethical grounding, and socio-legal continuity.

03

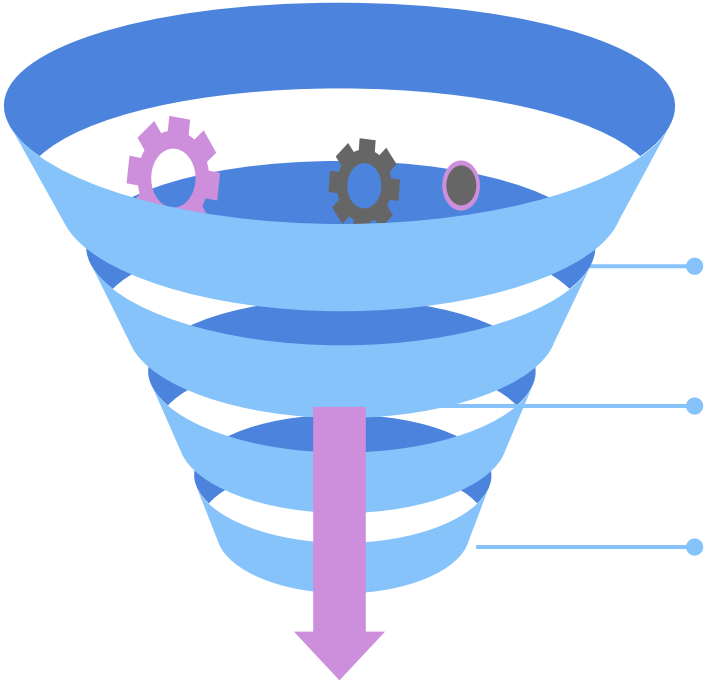
Trust Experience Design

Human trust is **felt experience**

Machine trust is **computational**

In the age of the agentic web, trust experience is both felt and computational

Trust Experience (TX) Design



1. **Trust Evidence** - Proof that an action really happened: on-chain logs, cryptographic attestations, explainable AI audit trails.

2. **Trust Primitives** - The low-level building blocks enabling evidence: decentralized identifiers (DIDs), verifiable credentials, multi-sig, zero-knowledge proofs.

3. **Trust Rituals / Experience Layer** - Human-facing practices and interfaces that let people feel trust: confirmations, ceremonies, ambient feedback, social endorsement.

Trust Experience (TX) Design

"Trust Experience (TX) Design" still a working definition (Stark, 2024)

TX = the total set of experiences that shape a person's expectations about a system's future behaviour - analogous to UX, but focused on trust formation; some dimensions include:

- **Epistemic verification**: Self-directed research (reading code, docs, economics) - *Reflexive Modernization*, Ulrich Beck (1994)
- **Social validation**: Trust is co-produced through networks of social relations (friends, experts, auditors, influencers) -Interpersonal trust (Luhmann, 1979) and communities of practice (Etienne Wenger, 1991)
- **Temporal reliability**: Historical performance (confidence, "Lindy effect" (Goldman, 1964) - longer something lasts, longer we expect it to last, precedent).
- **Collective legitimation**: mass social proofs, market as distributed judgement, crowd sentiment

It's a "**lifeworld**" problem - interacts with the taken-for-granted social, perceptual and political environment an artifact assumes in order to "make sense" (Wong et al. 2020)

Trust Experience (TX) Design

"**Hardness**" = Reliability you can bet on, cost to corrupt. A system is trusted when it would cost more to break it than anyone is willing (or able) to pay (Stark, 2024)

Protocols draw hardness from multiple layers—each reinforcing the other.

- **Physical**: atomic scarcity, energy cost, engineered durability (e.g. gold or rare earth)
- **Mathematical**: algorithmic unforgeability, cryptographic proofs, formal verification
- **Institutional**: law, regulation, standards, bureaucracy (e.g. legal system enforcing contracts, central banks guaranteeing currency)
- **Social**: collective belief, norms, reputational effects, Lindy time (handshake, handshake deals, reputation of a trusted local merchant)

Prototyping Protocols

Considerations	Description	Example
Tension	Design argument being tested	What conflicting values are at stake? E.g. Transparency vs. privacy
Level of fidelity	How real does the prototype need to be?	Sketch? Physical mock-up? Digital mock-up? Live pilot?
"Hardness" surface	Reliability levers and cost to break	Rarity? Math? Legal penalty? Social reputation?j
Second-order effects	Unintended outcomes	If everyone adopted it, what new behaviours, markets, or harms might emerge?
Adversarial moves	Ways an actor might deliberately exploit, bypass, or destabilise the protocol to gain advantage	Insider collusion? Regulatory arbitrage? Narrative attack? Brute force exploit?

Course Plan

Theory

Week 1
INTRO

Week 2
PROTOCOLS

Week 3
PROTOCOLS & TRUST

Week 4
TRUST & TECHNOLOGIES

Week 5
TRUST & TECHNOLOGIES

Research

Week 6
PROTOCOL WATCH

Week 7
PROTOCOL WATCH

Research

Week 8
PROTOCOL WATCH

Week 9
PROTOCOL WATCH

Studio: Prototype

Week 10
IDEATE / PROTOTYPE

Week 11
PROTOTYPE

Week 12
PROTOTYPE

Week 13
PROTOTYPE

Reflect

Week 14
FINAL PRESENTATION/ REFLECTION

04
Exercise

Design the “Traffic Lights” of the Agentic World” (Miro board)

If a green light lets us cross a busy street with confidence (along with many other protocols), what signals and rules will guide trust when intelligence is everywhere?

Imagine a near-future city where intelligence is distributed - in cars, buildings, household robots, street sensors, and AI agents that negotiate on your behalf.

Your task: Think about the trust protocols that keep everyday life safe and smooth.

Questions to consider:

- How do humans recognize and feel that an agent is trustworthy?
- What **technical primitives** guarantee correct behavior?
- What **governance and recourse mechanisms** handle failure or dispute?
- What **rituals or signals** make these assurances visible in everyday life?
- Failure and **evolution**: What happens when the protocol is attacked or norms shift? How can it adapt?

Pick a domain on the Miro board:

Autonomous Mobility

Design trust for roads full of self-driving cars and drones.

Domestic Robots

Kitchen bots, elder-care bots—how do we know they're safe?

AI Finance Agents

Wallet agents investing and negotiating 24/7.

Smart Buildings & Cities

Doors, elevators, energy grids that think for themselves.

Health & Bio Agents

AI triage doctors, gene-editing nano-bots.

Planetary Coordination

Swarms of climate-control satellites or ocean clean-up robots.

In your Miro board, work through the following. Feel free to use an LLM to design your scenario.

A. Scenario & Stakes	Describe your scenario in one sentence: 'In ____, an AI does ____ for ____; the risk is ____.' List key actors: e.g. Human(s), AI agent(s), third-party validators, regulators.
B. User Journey Steps	Map each step of the user journey: key action, data in/out, decision made.
C. What Could Go Wrong?	For each step, note potential risks: identity spoofing, data tampering, misalignment, lack of consent, operational failures.
D. Protocols Needed at Each Step	For each step, specify needed protocols and why (e.g., DIDs, zk-proofs, TEE attestations, on-chain logs).
E. Governance & Recourse	Describe oversight mechanisms: escalation ladder, dispute resolution, incentives or penalties.
F. Rituals & Signals	List visible or experiential trust cues: e.g. trust meter, verify button, ambient signals.
G. Evolution	Explain how trust will adapt over time: upgrade rules, reputation decay, post-mortems.

Assignment: Trust Experience Watch

Goal: Notice and analyze the *visible and invisible* trust protocols that make your daily activities possible.

Assignment (1 page or 3–5 slides)

1. **Pick two everyday situations** you experienced this week (e.g., crossing a busy street, online shopping, boarding a train, using a food-delivery app).
2. For each situation, **map the trust stack**:
    - **Physical / Material** (locks, traffic lights, packaging)
    - **Mathematical / Technical** (passwords, encryption, sensors)
    - **Institutional** (laws, warranties, company policies)
    - **Social / Ritual** (norms, greetings, social proof)

3. Briefly reflect on:
    - Where do you see **gaps or single points of failure**?
    - How might these protocols need to evolve as AI and autonomous agents become common?

Email me @: hr2703@nyu.edu for feedback