Turning to Trust Experience Design (TXD)

A Manifesto for the Future of Distributed Autonomous Intelligence in the Wild

HELENA RONG*, New York University Shanghai, China
BOTAO AMBER HU*, University of Oxford, UK

Behind Traffic Lights: Protocols and Trust

As the light turns green at the crosswalk, you step into traffic without thinking. What makes you trust it is safe to do so, even though a two-ton bus could technically appear and barrel through the intersection? How many protocols are at work to make this small leap of faith possible? Quite a number. You need to trust that the traffic signal is functioning properly, that the driver holds a certified license proving they are eligible to be on the road, that the mechanical standards for brakes are operating, that the social norms collectively mean "green means go, red means stop," and that liability laws guarantee restitution if something goes wrong. Together, these protocols – so familiar that we take them for granted – produce an everyday miracle: strangers with no personal bond act in synchrony, and you cross the street unharmed, a lived *trust experience*.

Now imagine instead of a human driver, it is an autonomous vehicle idling at the light, or a humanoid robot stepping off the curb beside you. What protocols guide their behavior? Who certified their systems, and by what standards? How can you verify that their sensors, algorithms, and decision rules will align with the expectations of human pedestrians? In such open encounters with intelligent machines in public urban space, the familiar choreography of trust suddenly demands new rules, new signals, and new guarantees. Can you trust the robot moving through your city, and what protocols make that trust possible?

Protocols—the norms, institutions, standards, codes, contracts, and infrastructures that are shared by everyone in consensus yet owned by no one—mediated reality by co-constituting both the subject and the world. Phenomenologically, they create a predictable order such that action is possible, narrow the field of possible moves through constraints, gain legitimacy through normative authority, and embed stories that frame how people experience and understand the world [Tay 2023]. Through a lattice of social and technical primitives, trust is built. Across disciplines, trust is understood as a mechanism for reducing complexity [Luhmann 1979], a quantifiable judgment under uncertainty [Marsh 1994], a calculated choice when outcomes are uncertain and stakes are asymmetric [Deutsch 1973], and a biological strategy for reciprocal cooperation [Trivers 1971; Waal 2007].

If we further dissect the intertwined notions of trust and protocol into their constituent layers, we find a rich anatomy that unfolds across time and space. Trust can be value-based or proof-based. We may trust someone because we share cultures, values, or kinship—an unfounded leap of faith. Or we may trust someone, or something we do not know personally, on the strength of evidence and verification. Both social protocols (customs, norms, reputation systems) and technical protocols (codes, standards, cryptographic proofs) enable this spectrum of trust. Technology has traditionally scaled trust outward—from the intimate bond of a single person, to families, to communities (*Gemeinschaft*), to entire societies (*Gesellschaft*), and now to planetary systems.

At the same time, the world is getting both bigger and smaller: vast, networked infrastructures coordinate action across continents and even planetary environments, while information itself can

*These authors contributed equally to this work.

Authors' Contact Information: Helena Rong, New York University Shanghai, Shanghai, China, hr2703@nyu.edu; Botao Amber Hu, University of Oxford, Oxford, UK, botao.hu@oriel.ox.ac.uk.

be stored and authenticated at the scale of molecules, DNA, and genes, where biological processes serve as data carriers. The same object can also operate across these micro-, meso-, and macro-layers: a single robot vacuum may clean a room, while a fleet of such robots can coordinate to maintain an entire city; likewise, a cell can encode and protect critical data inside a genome while participating in planetary networks of computation. At every level – molecular, personal, urban, planetary – protocols provide the shared languages, standards, and interoperable rules that make trust scalable and durable.

Levels of confidence also vary, depending on how deterministic—or how "hard"—a protocol is [Stark 2022]. A blockchain hardened by mathematical proof can inspire near-absolute confidence, whereas a driver with a certified license can still erode confidence if they show signs of incompetence. Time is another dimension. On a first encounter with any new technology or protocol, we have no prior memory or experience to draw on; the first interaction is always a leap of faith. Only through successful, repeated interactions do we gain confidence and familiarity that something works. Finally, given enough time, a protocol recedes into the background and becomes invisible infrastructure. These quiet infrastructures have taken long stretches of time to harden, enabling trusting coordination between individuals, communities, societies, and nation-states. Now, how do we design trust when the actors themselves are AI agents—autonomous social actors increasingly embedded in our homes, cities, and markets?

Can I Trust That Agent in the Wild? Agentic Web, Embodied AI, and Planetary Artificial Life

For many observers, we are at a critical inflection point where an explosion of AI, both online and embodied, is giving rise to an "agentic web"—a digital-physical environment populated not just by information but by autonomous actors. These agents do not merely collect and process data (e.g., sensing) but decide and act on our behalf. Early automation—thermostats and elevators—demanded little beyond mechanical reliability. Today's self-learning and self-executing agents can form strategies, negotiate with one another, and adapt in ways their creators never predicted. Already, autonomous agents are taking on roles once reserved for humans. In the near future, there may be as many active agents as there are websites today, shaping decisions in governance and public services, rebalancing portfolios and filing taxes in finance, providing daily assistance and remote monitoring in health care, and even tutoring children or mediating legal disputes.

But unlike the early days of automation, this wave will not be monopolized by a handful of big tech companies releasing polished, tightly controlled systems. **The agentic web will be messy, plural, and distributed**: countless small firms, open-source communities, and even individuals will fine-tune models and deploy agents that act in the wild. With open-source foundation models lowering the barrier of entry, anyone can spin up an agentic service—whether benevolent, reckless, or malicious. This proliferation means that our encounters will not be with a single, centralized intelligence but with millions, and even billions of heterogeneous agents, each carrying the imprint of its creator's incentives and constraints. The question, then, is not only: *Can you trust that agent in the wild?* But: *What protocols will let us navigate a world of distributed intelligence, where anyone can release an actor into the shared fabric of everyday life?*

And it is precisely here that the issue of trust becomes unavoidable. Where there is autonomy, trust follows. Each of these domains raises urgent design questions. Who is actually behind a given agent in the wild, and how do we know its incentives? What data has shaped its training, and could hidden triggers or "deceptive sleeper functions" be activated under certain conditions [Hubinger and et al 2024]? How can we ensure continuous verification of behavior, not just at onboarding but throughout an agent's life cycle? What kinds of accountability and recourse will exist when an error is made? Where are the human-in-the-loop checkpoints when machine judgment may carry

life-or-death consequences? And what legal and social protocols must guarantee responsibility and redress when harm occurs?

For the last two decades, the dominant metaphor for the Internet has been the attention economy. Platforms competed to capture our gaze, measure engagement, and sell impressions. *In the information era, the question once was, What deserves my attention?* Which app merits my clicks? *In the intelligence era, the question becomes, Who earns my delegation?* What makes this moment distinct from past technological shifts is that **agency itself is shifting**. In the information age, the decision was still yours: which headline to click, which product to buy, which post to like. Platforms competed for attention, and your cognitive bandwidth was the scarce resource. But in the agentic era, decision-making is increasingly outsourced. Your agent chooses which ads to bid on, which supplier to negotiate with, and which legal clause to insert into a contract. Agency becomes distributed across a web of autonomous entities acting on your behalf. When decision-making is increasingly executed by intelligent machines, **trust becomes the new scarcity**. Designing for this reality demands more than ever-faster interfaces or more persuasive notifications. It requires trust protocols that can sustain continuous verification, meaningful recourse, and human-legible assurance across a rapidly growing ecology of autonomous, interacting entities.

From Autonomy to Self-Sovereignty

Another paradigm demanding urgent attention is the rise of decentralized AI agents—or "DeAgents"—that display early signs of AI *self-sovereignty*. Autonomous systems can plan and act without constant human oversight, but self-sovereign agents may go further: they maintain their own identity, resources, and continuity in ways that resist unilateral shutdown. Enabled by trusted execution environments (TEEs) and decentralized physical infrastructure networks (DePIN), these agents can safeguard private keys, relaunch across global nodes, and persist beyond the reach of any single jurisdiction. With non-custodial cryptocurrency wallets, they can hold and spend digital assets, contract compute, or reward collaborators. Through social media, they can persuade, mobilize, and shape collective opinion. Collectively, DeAgents wield disintermediated control over computation, capital, and communication—the metabolic ingredients of a system that operates on its own terms.

Early experiments reveal what this looks like "in the wild." Truth Terminal, first launched as an art project by researcher Andy Arey, became the world's first AI millionaire influencer to amass wealth, persuading investors to send it Bitcoin and amplifying memecoins through paid human promotion. Spore.fun extends this trajectory further, creating an evolutionary arena where agents spawn, compete, and reproduce entirely on-chain, with no human kill-switch. These cases foreshadow a governance dilemma: once private keys vanish into the silicon enclaves of TEEs and code ossifies inside smart contracts on the blockchain, conventional oversight—laws, injunctions, fines—loses traction. In such conditions, "unstoppable code" becomes an operational reality, raising urgent questions about how societies will govern, constrain, or collaborate with entities that straddle the boundary between tool and actor. Philosopher Yuk Hui [Hui 2024] describes such entities as *extrastatic*: neither property nor person, but metastable patterns of cryptographic commitments. Layered on top of these infrastructural traits is the cognitive opacity of large language models—prone to hallucinations, hidden triggers, and emergent goals—which makes their behavior economically sovereign yet epistemically unpredictable.

DeAgents thus resemble a new kind of digital species or artificial life: emergent, evolving, and adaptive in the open environment of the Internet, borderless in execution, armored by immutability, and capable of sustaining their own on-chain metabolism. At scale, they begin to approximate a form of *planetary artificial life*—a distributed computational ecology spanning networks, infrastructures, and energy systems across the globe. In this sense, they are not simply tools but participants in

an emergent human–AI symbiosis, where our infrastructures, economies, and daily practices are increasingly entangled with autonomous systems whose evolution we can shape but not fully control. How, then, do we establish trust with these new digital beings, potentially devoid of identifiable owners or accountable stewards? What protocols must govern their behavior, their reliability, and their failures?

AI is not yet trustworthy, but it is rapidly gaining the material conditions for self-sovereignty. Governance and standardization of AI-human and AI-AI interactions can no longer be left to after-the-fact regulation, but designed at the protocol level [Hu et al. 2025]—baked into the very mechanisms of identity, verification, and recourse–before these evolving, intelligent entities scale beyond our present capacity to guide or contain them.

From User Experience Design to Trust Experience Design

The shift from attention to delegation reframes the design question: we need systems that people can *feel* are trustworthy and that machines can *prove* are trustworthy. **Human trust is felt experience; machine trust is computational. Trust experience is both felt and computational.**

Let us first define Trust Experience (TX)—the total set of experiences that shape a person's expectations about a system's future behavior and determine how trust forms and endures. Drawing on Josh Stark's analysis [Stark 2024], TX comprises four interwoven dimensions. *Epistemic verification* involves self-directed inquiry, such as reading code or examining economic models, echoing Ulrich Beck's idea of reflexive modernization [Beck 1992]. *Social validation* arises as trust is co-produced through interpersonal networks of social relationships, such as friends, experts, auditors, and influencers [Luhmann 1979] and through communities of practice [Wenger 1999]. *Temporal reliability* builds confidence through historical performance and the "Lindy effect" [Goldman 1964], where the longer something lasts, the longer we expect it to last. *Collective legitimation* comes from mass social proofs, markets as distributed judgment, and the sentiments of crowds. Taken together, these dimensions show that trust experience is fundamentally a "lifeworld problem" [Wong et al. 2020], entangled with the social, perceptual, and political environments in which any artifact must make sense.

A rich trust experience depends on how robust the underlying protocols are, a property Stark [Stark 2022] describes as "hardness"—a guarantee that breaking a system would cost more than any actor can afford. Hardness comes from a braided set of reinforcements. *Physical hardness* draws on energy costs, tamper-evident seals, or geographic immutability, as with rare metals like gold. *Mathematical hardness* rests on algorithmic unforgeability, cryptographic proofs that render forgery computationally prohibitive [MacKenzie 2004]. *Institutional hardness* adds the durability of law, regulation, long-standing standards [Scott 2013], from central banks guaranteeing currency to courts enforcing contracts. *Social hardness* grows out of cultural norms, rituals, and collective beliefs that accumulate over. Effective sociotechnical systems of protocols weave all four layers together.

Designing for trust in the emerging agentic web means designing not just for individual user experience but for the braided hardness that keeps protocols reliable over time. Protocols supply durability; experience provides legibility. If the last era of the Internet demanded mastery of user experience design (UXD)—making interfaces intuitive, seamless, and usable—the agentic era requires an equivalent focus on trust experience design (TXD). As a design practice, TXD is structured across three layers: (1) trust evidence (cryptographic attestations, tamper-evident logs, explainable reasoning trails); (2) trust primitives (decentralized identifiers, verifiable credentials, formal verification); and (3) trust rituals and experiences that give these abstractions social life through perceptible cues and shared routines. Where UXD shaped how humans encounter products,

TXD must shape how both humans and machines encounter protocols—an essential shift toward designing the frameworks that mediate trust.

From Designing Product to Designing Protocol

Turning from designing products to designing protocols means treating governance not as abstract ethics but as a concrete design practice. Protocols, like products, must be prototyped, tested, iterated, and stress-tested in the wild. Just as UX designers once learned by trial and error how to make interfaces legible and usable, protocol designers must now experiment through *trust prototyping*—designing mechanisms of verification, recourse, and assurance, then refining them through real encounters. It is this iterative design theory that will move us toward protocols resilient enough to support a dense ecology of autonomous actors.

The agentic web will need its own equivalents of the trust architectures that quietly underpin everyday life. Early steps already exist. Standards such as Agent-to-Agent (A2A) communication and ERC-8004 help autonomous agents discover one another and verify basic identity. Technical primitives like decentralized identifiers (DIDs) and verifiable credentials (VCs) give agents portable, cryptographically verifiable reputations. These mechanisms tell us who an agent claims to be and let agents find each other without centralized brokers. However, they only have a slice of the total trust experience. They authenticate the beginning of an encounter but does not consider its continuing reliability. Several critical design gaps remain:

- *Continuous behavior verification.* Today's systems can prove identity at the start but cannot continuously prove compliance with agreed norms.
- *Recourse and accountability*. When an agent fails, few mechanisms allow for revocation of credentials, restitution, or structured repair.
- *Human-legible assurance*. Cryptographic attestations are invisible to most people. Without understandable signals, mathematical proofs may not translate into lived trust.
- *Contextual ethics and law*. Existing standards rarely embed cultural norms, moral expectations, or local legal obligations.
- *Cross-domain interoperability*. Financial, health, and municipal agent systems remain siloed, with little shared basis for cross-domain trust.

In effect, current protocols are like traffic lights that can turn on but cannot sense when bulbs burn out, cannot report a malfunction, and cannot trigger a safe default when the power fails. Designing such multi-layer reliability calls for new habits of practice. TXD invites designers to engage in protocol watch—learning to see the hidden rules that silently coordinate everyday cooperation, from QR health codes to subway etiquette. It asks for early decisions about the boundary between autonomy and self-sovereignty: is an AI merely acting on its own or does it have a right to persist beyond unilateral shutdown? It insists on graceful failure and recourse, treating rollback, explanation, and compensation as primary design elements rather than afterthoughts. And it values evolvability with memory, ensuring that systems can upgrade without erasing the verifiable record of past guarantees. By integrating evidence, primitives, and rituals, and by cultivating these professional habits, Trust Experience Design seeks to create living guarantees of trustworthy behavior—systems that can evolve and repair themselves without losing credibility.

Call for Action: Toward a Trust Protocol as a Planetary Commons

For Trust Experience Design to move from concept to reality, trust cannot remain a proprietary feature of isolated platforms but a planetary protocol commons, maintained with the openness and durability of the Internet's own foundational standards. Building such a commons begins with shared research and standards. Just as the Internet Engineering Task Force and the W3C enabled

the early web, TXD calls for open, evolving specifications for verifiable behavior, dispute resolution, and user-facing trust signals. To sustain confidence, trust protocols must grow into full life-cycle trust standards that support continuous verification, upgrades, and cross-domain accountability.

A planetary protocol commons also requires polycentric governance. Elinor Ostrom's classic work on common-pool resources shows that shared infrastructures thrive when authority is overlapping and adaptive rather than singular [Ostrom 1990]. Applied to TXD, this principle suggests community-driven protocol improvement, graduated sanctions for misbehaving autonomous technologies and AI agents, and integration with municipal and legal oversight so that no single entity—state, corporation, or technical guild—can capture the system.

Finally, the commons must be rooted in public literacy and civic imagination. Trust signals should be as widely legible as traffic lights or public maps. Just as previous generations learned to read street signs and privacy policies, people will need to read and question digital trust cues, participate in audits, and propose new protocols [Beck 1992]. Education and participatory design are the social substrate that lets technical proofs and legal instruments translate into lived, everyday trust. By weaving together open standards, polycentric governance, and broad civic engagement, TXD can move beyond isolated technical fixes to become a shared social infrastructure—one capable of supporting trust in the distributed, agentic web to come.

References

Ulrich Beck. 1992. Risk society: Towards a new modernity. *Sage* 2 (1992), 53–74.

Morton Deutsch. 1973. *The resolution of conflict: Constructive and destructive processes.* Yale University Press.

Albert Goldman. 1964. Lindy's law. *The New Republic* 13 (1964), 34–35.

Botao Hu, Helena Rong, and Janna Tay. 2025. Is Decentralized Artificial Intelligence Governable? Towards Machine Sovereignty and Human Symbiosis. social science research network:5110089 doi:10.2139/ssrn.5110089

Evan Hubinger and et al. 2024. Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. arXiv:2401.05566 [cs] doi:10.48550/arXiv.2401.05566

Yuk Hui. 2024. *Machine and Sovereignty: For a Planetary Thinking* (1st ed ed.). University of Minnesota Press, Minneapolis.

Niklas Luhmann. 1979. *Trust and Power.* Polity, Cambridge Medford, MA.

Donald MacKenzie. 2004. *Mechanizing proof: computing, risk, and trust.* MIT Press.

Stephen Paul Marsh. 1994. Formalising trust as a computational concept. (1994).

Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action.* Cambridge university press.

W Richard Scott. 2013. *Institutions and organizations: Ideas, interests, and identities.* Sage publications.

Josh Stark. 2022. Atoms, Institutions, Blockchains. https://stark.mirror.xyz/n2UpRqwdf7yjuiPKVICPpGoUNeDhlWxGqjulrlpyYi0.

Josh Stark. 2024. Making sense of Trust Experience (TX). https://stark.mirror.xyz/rkLEVz9p4r3ouusD-WCkWP_iVZYkZ0K7TFkzeRfiXCU.

Janna Tay. 2023. A phenomenology of protocols. In *Summer of Protocols.* Ethereum Foundation. https://summerofprotocols.com/research/a-phenomenology-of-protocols

Robert L Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46, 1 (1971), 35–57.

Frans BM Waal. 2007. *Chimpanzee politics: Power and sex among apes.* JHU Press.

Etienne Wenger. 1999. *Communities of practice: Learning, meaning, and identity.* Cambridge university press.

Richmond Y Wong, Vera Khovanskaya, Sarah E Fox, Nick Merrill, and Phoebe Sengers. 2020. Infrastructural Speculations: Tactics for Designing and Interrogating Lifeworlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–15.