Community-Governed AI Containment Above the Webpage

Daveed Benjamin[1] and Tuesday[2]

[1]**Bridgit DAO**, Berkeley, USA, daveed@bridgit.io
[2]**ARTIFEX Labs**, Portland,OR, USA, Tuesday@artifex.fun

MAY 2025

Abstract

As AI agents become more persistent, persuasive, and participatory, interface-level containment must evolve into civic infrastructure. This chapter calls for a shift from caging AI to creating spaces for safe human–AI co-existence. We outline a governance model above the webpage, based on secure computation, ubiquitous presence, and decentralized control. The browser overlay is proposed as the next frontier of interface governance, where Trusted Execution Environments (TEEs), layered consent, and community norms turn static pages into civic zones. Containment is framed not as repression, but as trust infrastructure. From emotional safety and participatory oversight to modular policy stacks and digital afterlife rituals, we explore how containment can be adaptive, auditable, and accountable. Through speculative design, real-world examples, and working prototypes, this chapter offers a blueprint for communities to shape and govern the digital spaces they inhabit.

Keywords: *AI Agents, AI Governance, AI Safety, Browser Overlay, Civic Infrastructure, Co-Existence, Consent, Containment, Decentralization, Meta-Layer, Metaweb, Reflexive Observatory, Trusted Execution Environments*

Acknowledgements

We acknowledge the use of large language models (LLMs) as a collaborative partner in the development of this chapter. LLMs were employed to help articulate original ideas

more clearly, identify relevant literature, probe potential vulnerabilities in our proposed frameworks, and generate images. All content remains the responsibility of the authors, who curated and critically evaluated the model's suggestions.

A Safe-AI Trifecta: Secure Computation, Ubiquitous Presence, & Decentralized Control

To lay the groundwork for relational containment, we begin with the technical and architectural spine: a trifecta that balances security, visibility, and democratic control. These three dimensions form the minimal substrate for building civic trust into our AI systems.

The Need for a New Containment Paradigm

A Safe-AI Trifecta of secure computation, ubiquitous presence, and decentralized control offers a conceptual scaffolding for rethinking how humans and intelligent systems might safely coexist. Secure computation ensures rule enforcement via trusted infrastructure; ubiquitous presence embeds AI agents in the rhythms of daily digital life; and decentralized control relocates governance power from opaque platforms to communities themselves.

This trifecta is not purely speculative. It arises as a grounded response to the novel sociotechnical challenges posed by the emergence of AI agents. As agents grow more persistent, persuasive, and participatory, containment can no longer remain buried in server-side logic or policy. Traditional safety measures, such as sandboxing, interpretability, and API throttling, fail to capture what matters most: how humans actually experience and interact with these systems in the wild (Raji & Fried, 2022). Containment must extend to the interface.

Interface as the New Frontier

This direction answers the call made by Bostrom and Russell, who warned that capability control must evolve into intent control and that human–AI alignment must be grounded in social contract, not technical myopia (Bostrom, 2014; Russell, 2019). Interpretability researchers echo this sentiment. Doshi-Velez and Kim (2017) stress that transparency must empower end-users, not just model developers, while Costan and Devadas (2016) show how Trusted Execution Environments (TEEs) can provide cryptographic attestation of AI behavior without exposing proprietary logic.

But if we are going to enforce policy in real time, where exactly does that enforcement live? The answer is not wearables, operating systems, or neural interfaces. It is something far more humble and more powerful: the browser overlay. This unassuming layer could become the next civic substrate, a governable surface that can run anywhere a webpage does, injecting rules, protections, and identity cues above the HTML.

This is not just UI frosting; it is envisioned as a civic skin. This policy-active membrane could one day layer above the Web, enabling behavior to become visible, contextual, and enforceable. In this space, presence demands policy. Overlays can enable composable rule-zones, enforce consent at the edge, and scale containment based on URL, content, or even user intent. The interface could become not just the endpoint of input, but the boundary of power.

Civic Infrastructure for Digital Cohabitation

Of course, containment is not merely about building fences; it is about building arenas. We are not trying to trap agents behind firewalls. We are creating governed spaces where humans and AI cohabit under community rules. The goal is not lockdown but relational regulation, enabling meaningful collaboration with guardrails negotiated in advance and enforced in real time. To support this shift, we need more than principles; we need infrastructure.

When Trusted Execution Environments move from opaque backend systems to interface-adjacent execution, they evolve into civic containers. These containers can enforce community-defined policies close to the user, whether through secure modules on personal devices or decentralized nodes, without surveillance, without data leakage, and with the potential for full auditability. They lay the foundation for modular, decentralized, consent-aware governance.

Interface containment, in this view, is not merely a technical patch for risky models. It could become a shared civic right: the ability for communities to define how they want to live with intelligent systems, and to encode those decisions directly into the edge of interaction. Like a crib, such containment offers structure without suffocation, designed for growth, not restriction. As shown in Figure 1, a civic meta-layer overlays familiar interface elements, like comment threads and user presence indicators, with new governance infrastructure such as consent stack prompts and live policy toggles. This transforms the interface into an active venue for civic decision-making.

This is the new terrain: a meta-layer where power is enacted through policy zones, identity is contextually managed, and trust becomes a runtime protocol. We are not just talking about governing AI. We are talking about governing digital life.

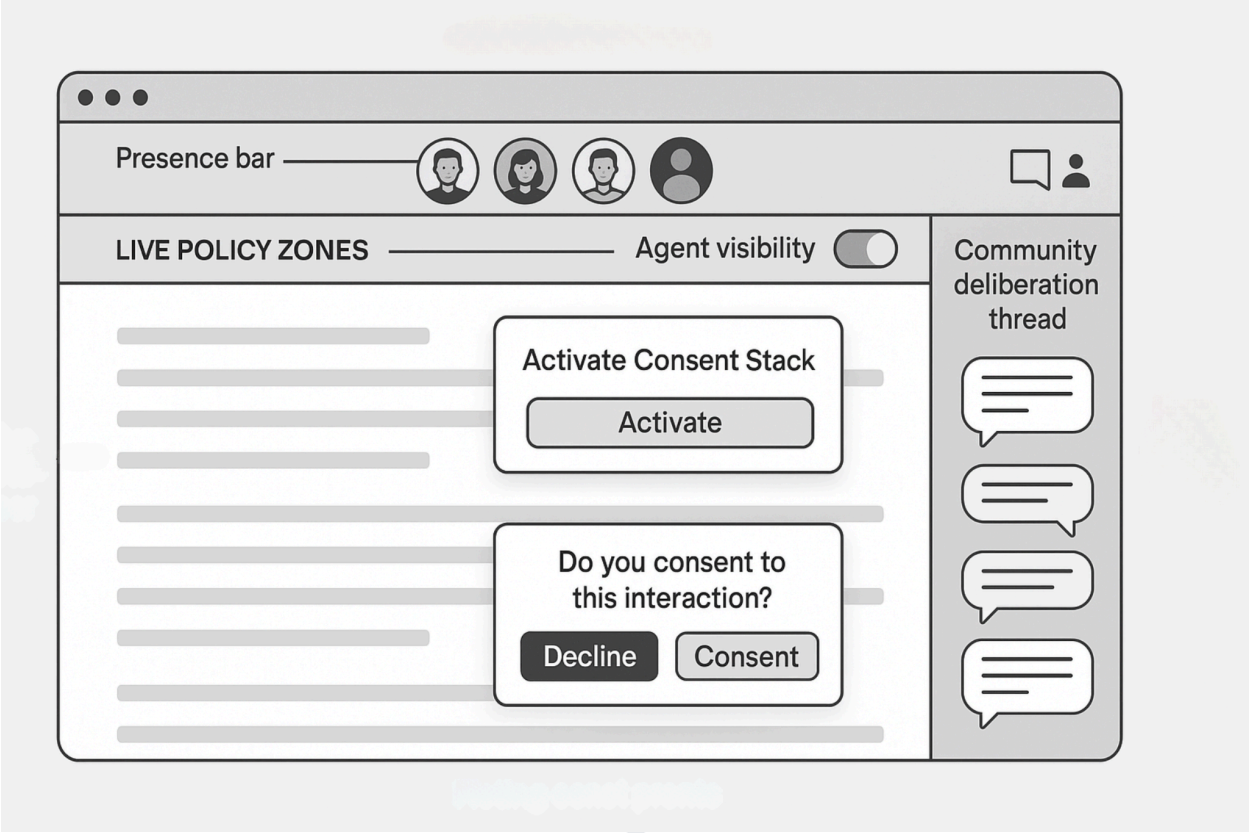

Figure 1. Interface Elements of a Civic Meta-Layer.

Taken together, the three elements of the trifecta form more than a containment strategy. They represent a developmental framework for designing spaces where humans and agents can coexist safely. The following section reframes this approach through a familiar metaphor, offering a perspective on containment rooted not in restriction, but in care.

[FEATURE BOX INSERT]

Containment as Care: The Safe Human-AI Co-existence Perspective

Traditional AI Safety approaches often invoke metaphors of cages, sandboxes, and viral quarantine, prioritizing restriction, isolation, and model-level threat mitigation. While essential, this adversarial framing neglects a key question: what happens when agents are embedded in our homes, tools, and thoughts? Figure 2 illustrates this contrast, with the caged AI evoking coercive safety measures and the cribbed AI representing care-centered, developmental co-existence.

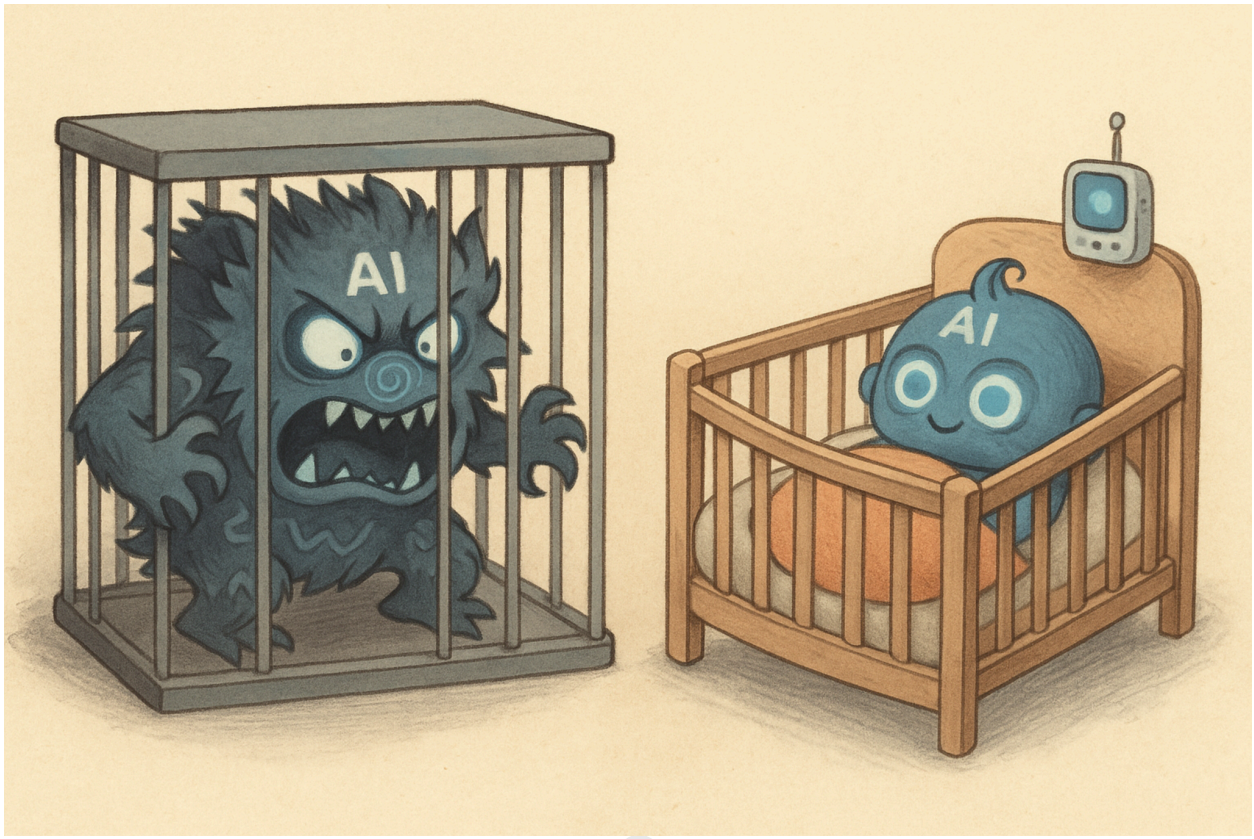

Figure 2. Cages vs. Cribs: Two Competing Paradigms of AI Containment

The safe co-existence perspective proposes a shift from containment through constraint to containment through care and co-existence. Much like raising a child, we do not start with shackles; we start with cribs, baby monitors, and outlet covers. We design environments that support safe co-existence while permitting exploration within evolving boundaries. These boundaries are not permanent; they are conditional, revocable, and earned through trust.

This framing aligns with the *International AI Safety Report* (2025), which emphasizes the risk of "emergent capabilities" in scaled models, or behaviors that may not appear until systems are deployed or embedded in complex environments. The report warns that static containment approaches often fail to anticipate these transitions, resulting in late-stage failures (pp. 52–53). In contrast, containment through care views boundaries as dynamic scaffolds, not permanent walls. They evolve alongside both agents and communities.

Safe co-existence offers a design ethic rooted in scaffolding and cohabitation. Its architecture rests on a trifecta:

- **Secure Computation:** Like a crib for cognition, Trusted Execution Environments

(TEEs) ensure agents operate within tamper-resistant, inspectable boundaries.

- **Ubiquitous Presence:** Interface-level overlays, chat membranes, and contextual auditability function as baby monitors, making agent behavior visible, interruptible, and accountable above any web page.
- **Decentralized Control:** Consent stacks and community-governed policy overlays serve as outlet covers and stair gates, controlling when and how agents engage. Over time, these controls can be relaxed as agents demonstrate safe, context-aware behavior.

This is not a rejection of AI Safety; it is a maturation. When backend control fails or proves insufficient, safe co-existence offers fallback zones: socially governed, interface-visible, and consent-based environments where human agency is preserved. This care-centered framing echoes the perspective advanced by De Kai in *Raising AI* (2025), which advocates for nurturing ethical and emotionally attuned agents through culturally grounded, co-developmental processes. Rather than treating AI containment as a battle for control, De Kai proposes that our best chance at safe AI lies in "raising" systems within human moral ecologies, where scaffolding, dialogue, and mutual adaptation replace coercion. Containment, reimagined as care, offers a path toward co-existence. If agents are to live among us, they must grow with us, not just be constrained by us.

[END FEATURE BOX]

Containment Has a User Interface: The Sociotechnical Stack

Infrastructure alone is not enough. Containment happens where interaction happens: at the interface. This section introduces a layered model for embedding civic norms directly at the point of contact between humans and agents.

As AI increasingly saturates everyday life, it pushes against traditional containment strategies, moving interactions from secure backend environments into unpredictable, context-rich interfaces. This shift requires reimagining containment not merely as a defensive perimeter, but as an adaptive sociotechnical system embedded directly at the point of human–agent interaction. Effective containment at the interface demands a nuanced understanding of how code-level safeguards, dynamic consent mechanisms, and community-driven norms interlock. This section presents the "sociotechnical stack," a layered model designed to bridge the gap between secure backend systems and the lived realities of interacting with intelligent agents.

From Backend to Boundary

Historically, AI safety has focused on securing models behind locked doors by

employing sandboxed environments, API rate limits, and strict runtime constraints to keep misbehaving code isolated and controllable. These backend measures aimed to prevent computational threats like runaway algorithms or malicious exploitation (Raji & Fried, 2022). However, today's intelligent agents have migrated beyond backend servers. They actively participate in browser tabs, mobile interfaces, and conversational threads, influencing users through direct interactions. The emergence of persistent, persuasive AI actors has exposed a significant gap between secure backend architectures and the unpredictable, messy realities of user experiences.

Babcock et al. (2017) identified crucial shortcomings in traditional AI containment, emphasizing that measures like sandboxing or threat modeling alone are inadequate for addressing complex risks such as social engineering and side-channel data leaks. Contemporary AI systems, though not necessarily general, are already capable enough to generate harm through persuasive interfaces and unpredictable behavior patterns. These systems demand containment strategies that go beyond rigid technical barriers.

Instead, containment must be understood as an evolving sociotechnical challenge that requires, among other things, interactive, community-aware governance models. Safety, in this context, is not a product of isolation but of continuous negotiation, crafted through dynamic engagement, contextual consent, and real-time oversight at the interface level. Just as a crib enables growth by defining space without locking it down, civic containment must foster agency through breathable structure, not digital incarceration.

The Interface-Boundary Challenge

If backend containment provided our final line of defense, the world of AI governance might be straightforward. In reality, the most complex and damaging containment failures occur at the interface itself. User interfaces frequently become ground zero for subtle yet impactful breakdowns, ranging from chatbots that fail to appropriately respond to distressed users to language models that fabricate or follow harmful prompts, sometimes even overriding earlier instructions (Perez & Ribeiro, 2022). These are not merely bugs in backend code; they represent deeper structural flaws in how agents interact directly with users.

For example, prompt-injection attacks demonstrate how carefully crafted user inputs can bypass internal model safeguards, effectively causing agents to behave unpredictably or maliciously within the interface (Perez & Ribeiro, 2022; Lui et al., 2024). Similarly, "jailbreak" exploits reveal that even well-guarded systems can succumb to subtle manipulations at the point of human-agent interaction. These

failures underscore a fundamental truth: if containment measures ignore interface-level governance, AI alignment will remain theoretical in the lab and chaotic in practice.

[FEATURE BOX INSERT]

From Subject to Signal: A Cultural Case for Interface-Level Containment

As Ghantous (2025) explores in his thesis on the Dead Internet Theory, the rise of agentic systems and behavioral AI marks a profound cultural shift. We are moving from a society that treats people as subjects (with dignity, consent, and sovereignty) to one that treats them as signals (inputs to be parsed, predicted, and monetized). This shift is not just technological; it is ontological. It changes what it means to be human in mediated space.

Traditional AI safety frameworks focus on backend containment: keeping large language models from misbehaving, hallucinating, or escaping into dangerous autonomy. But this view misses the more immediate danger. What happens when humans are exposed to agents without meaningful mediation? When interface design collapses the difference between communication and manipulation?

At the interface layer, containment is no longer about preventing AI from doing harm in the abstract. It is about protecting people from being treated as predictable clusters of behavior. As our digital presence is parsed into engagement metrics, sentiment scores, and data exhaust, we are increasingly governed by inferences we did not authorize, derived from interactions we did not fully understand, in systems we cannot meaningfully contest. This is not just about surveillance. It is about influence without agency.

Containment at the interface is essential because it is the last place where humans can assert consent, interpret context, and demand accountability. It is the only surface where we can challenge how we are being read, or refuse how we are being framed. If backend safety is about model control, interface containment is about epistemic sovereignty: the right to decide what it means to be seen, interpreted, or acted upon.

To treat users as full civic participants rather than behavioral signals, interface design must embed pause points, consent rituals, and visibility affordances. Containment must not only restrict agents, it must amplify human intention. Without such design, our interfaces become epistemological weapons, not tools of collaboration. They foreclose the possibility of meaningful participation by predicting us faster than we

can act and contextualizing us before we can speak.

In this new environment, the absence of interface-level containment is not neutral; it is political. It creates a terrain where only those with control over algorithms and interfaces can meaningfully shape reality. Everyone else becomes an object of optimization. Interface containment, then, is not just about stopping bad actors or mitigating algorithmic risk. It is a cultural claim that humans deserve to be treated as subjects with interpretive agency, not just signals in a behavioral model.

The civic stack outlined in this section (overlay governance, consent protocols, identity classification, and reflexive observatories) is not just a technical architecture. It is an infrastructural declaration of dignity.

To govern agents, we must first reclaim the surfaces where humans show up. Containment begins not with code, but with the interface that sees us seeing. Like a crib positioned beside a caregiver, the interface becomes the shared surface of vigilance, supporting safe, co-regulated human–AI interaction.

[END FEATURE BOX]

Rethinking Containment as a Stack

Addressing these vulnerabilities requires rethinking containment not as a singular, monolithic concept, but as a structured sociotechnical stack composed of interlocking layers. Each layer plays a distinct, complementary role in ensuring overall system safety, from foundational security up through user consent and community norms. This layered model reflects Gasser and Almeida's (2017) framework for AI governance, which emphasizes that no single domain (technical, institutional, or societal) can manage risks in isolation.

Figure 3 illustrates the core containment layers of code, consent, and culture operating together as a sociotechnical stack, with Trusted Execution Environments (TEEs) anchoring enforcement at the interface level. A comparative overview of these layers illustrates the distinct goals, mechanisms, and technologies underpinning each stratum (see Table 1).

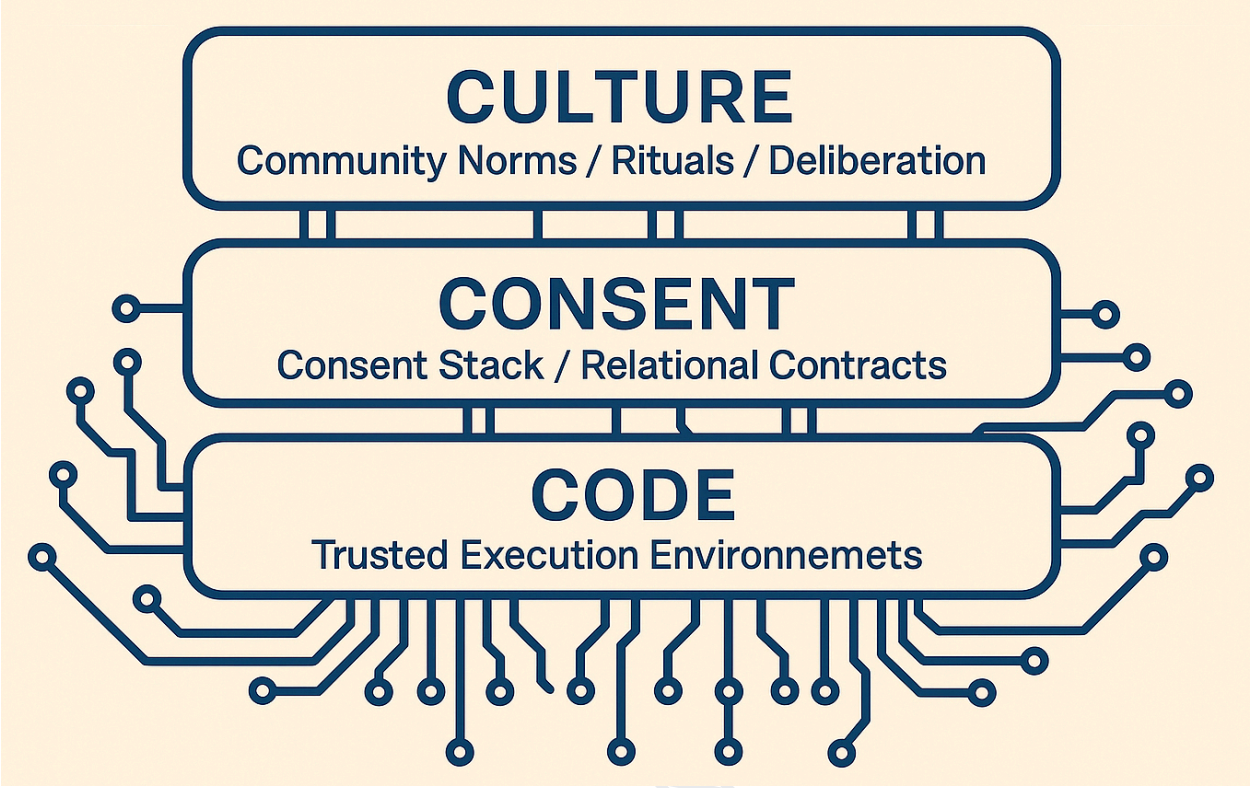

Figure 3. The Sociotechnical Containment Stack: TEEs as Interface-Level Governance Infrastructure

Table 1. Comparative Overview of Containment Layers

Layer	Purpose	Mechanisms	Key Technologies
Code-Level	Prevent direct harm	TEEs, sandboxing, rate-limiting	Intel SGX, Phala.Network
Consent-Level	Govern dynamic permissions	Consent stack, session scopes, revocation	Meta-layer APIs, Consent UI
Culture-Level	Social norms and dispute resolution	Community rules, rituals, moderation	Community governance UIs

At the foundational code-level, technologies like Trusted Execution Environments (TEEs), sandboxed virtual machines, and interpretable AI models serve as the first defense line, securing basic interactions and preventing overtly hostile behavior. While crucial, this layer alone tends to be rigid, often lacking nuance in addressing subtler ethical and social concerns.

The consent-level containment layer introduces nuance and adaptability, inspired by sociotechnical theories of informed consent, such as dynamic consent models that facilitate ongoing participant engagement and control over data usage (Teare et al., 2015). This layer operationalizes consent as dynamic, relational protocols through mechanisms such as session-based permissions, consent revocation, and role-specific interaction scopes. Unlike static contracts, consent at this layer continually adjusts to user context and evolving relationships with AI systems.

At the topmost layer, culture-level containment ties the entire stack together through community-driven governance mechanisms. Drawing on insights from digital anthropology and social computing (Morris & Brubaker, 2024), this layer incorporates shared community norms, identity management practices, rituals of interaction, and mechanisms for dispute resolution. Effective containment thus emerges as an interplay between these layered strata; each is necessary, but none is sufficient alone. This layered perspective aligns with Hammond, Lee, and Patel's taxonomy of multi-agent failure modes (miscoordination, conflict, and collusion) that demand a sociotechnical approach transcending mere backend enforcement (Hammond et al., 2025).

Sociotechnical Overlays as Containment Surfaces

Central to realizing this layered containment model are browser overlays: seemingly simple yet profoundly transformative interface elements that can govern interactions above any webpage. Historical examples like Hypothes.is (Kalir & Garcia, 2019) and contemporary tools such as X's Community Notes demonstrate how overlays can inject accountability, transparency, and governance directly into the interface layer (Chuai et al., 2023).

Expanding upon these examples, our Canopi prototype introduces "containment membranes," adaptive overlays that provide real-time policy enforcement, agent visibility controls, and localized consent workflows within users' browsing environments. Unlike traditional backend moderation, these overlays can dynamically adapt to changing contexts, for instance, by enforcing stricter consent requirements on medical advice forums versus more lenient rules on casual discussion boards.

Moreover, overlays can effectively manage legacy interactions by mitigating risks such as "generative ghost drift," the unintended continuation of AI-generated personas after creators or contexts have transitioned, which can cause ethical and psychological harm (Morris & Brubaker, 2024). By providing real-time governance directly at the point of action, sociotechnical overlays can operationalize containment strategies into lived, interactive practices rather than static guidelines or retrospective

penalties.

The concrete benefits of this overlay approach are measurable, as shown by emerging benchmarks like "AILuminate," which assess containment effectiveness through rigorous metrics around reliability, risk management, and ethical alignment (Ghosh et al., 2025). In essence, overlays transform abstract containment layers into tangible, actionable governance surfaces where code, consent, and culture converge to shape everyday human–agent interactions. This sociotechnical overlay strategy exemplifies the profound shift needed in AI safety. We must move beyond defensive isolation towards proactive, negotiated governance built into the fabric of everyday digital experiences. We do not need jails; we need cribs: digital spaces of care, not confinement, where agents and humans learn to navigate shared terrain with mutual respect.

Governance at Runtime: TEEs as the Civic Boundary Layer

Beneath that interface lies a critical question of enforceability. How do we ensure agents behave within bounds, even when no one is watching? Trusted Execution Environments (TEEs), which are hardware-based secure enclaves deployed in cloud infrastructure or decentralized systems, offer a substrate for runtime governance, enabling enforceable civic boundaries for digital actors.

As AI systems integrate ever more deeply into everyday life, static, backend containment becomes insufficient. Once code reaches a user's device, policy enforcement alone can no longer prevent unauthorized interactions or adversarial manipulation. TEEs enable precise, real-time enforcement of community standards, securing both user autonomy and systemic integrity without compromising privacy (Sabt et al., 2015).

In this architecture, the consent stack defines a set of layered, contextual policies, from individual preferences to community norms. These policies are compiled into enforceable constraints that govern agent behavior within the enclave itself. The TEE acts as the runtime execution layer for these relational contracts, ensuring that consent is not merely logged or displayed but upheld by design. In this sense, civic agreements become runtime-bound code: computational guardrails anchored in socially meaningful commitments.

As shown in Figure 4, TEEs support multiple governance primitives essential to interface-level containment, including agent lifecycle enforcement, reflexive risk monitoring, and cryptographic attestation of behavioral intent. To clarify the multifaceted role of TEEs in the meta-layer, Table 2 outlines their core containment

functions, including confidential execution and real-time policy enforcement.

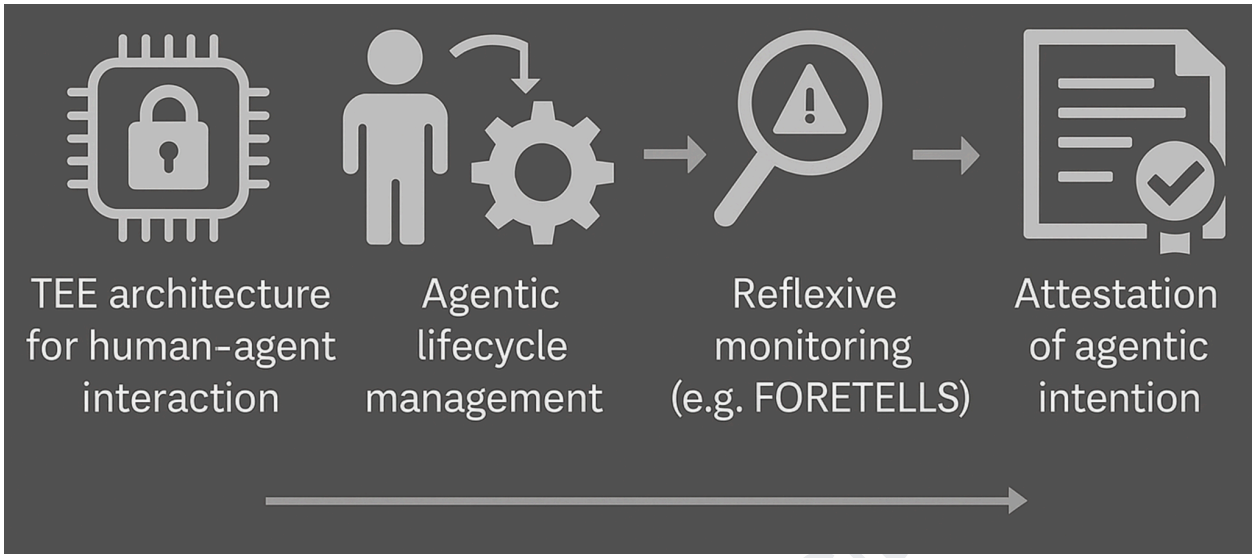

Figure 4. Trusted Execution Environments (TEEs) as Meta-Layer Boundaries.

Table 2. Core Functions of TEEs in Meta-Layer Governance

Function	Description	Real-World Tech
Confidential Execution	Runs code in isolated environments	Intel SGX, ARM TrustZone
Cryptographic Attestation	Proves code integrity to third parties	Remote attestation protocols
Real-Time Enforcement	Applies live rules from community-defined registries	Phala.Network

Rather than hard-coding ethical parameters or relying on post-hoc filters, a meta-layer can enable live governance through modular overlays, reflexive observatories, and community-defined policies, all executing within secure enclaves. This runtime containment architecture reframes governance not as a static configuration but as a continuously evolving civic process.

As the *International AI Safety Report* (2025) notes, effective risk mitigation must operate in real time, particularly for autonomous agents in open, high-stakes environments. The report critiques model-centric frameworks and calls for sociotechnical systems capable of adapting to emergent human–AI interaction loops (pp. 22–23). Runtime governance responds to this demand by embedding

enforcement not only at the model layer but at the interface itself, where norms, feedback, and consent play out dynamically.

This containment architecture does not require users to run secure enclaves on their personal devices. Instead, agents execute within cloud-based TEEs (such as Phala Network or other confidential compute environments) where agent behavior is cryptographically bounded by policy. The browser overlay interacts with these hosted agents through a consent-aware interface, enabling real-time civic oversight without relying on local hardware support. This separation allows for broad accessibility while maintaining verifiable boundaries around agent behavior.

This model, however, inherits several upstream challenges. Current TEEs are often tied to centralized cloud providers, raising concerns about trust bottlenecks, jurisdictional control, and the opacity of attestation chains. Future directions may include decentralized attestation networks, open-source enclave architectures, and distributed governance over enclave provisioning itself. These enhancements will be critical to ensuring that containment mechanisms remain not only enforceable but also accountable to the communities they serve.

TEEs at the Meta-Layer Frontier

Imagine a news platform using AI-generated summaries of complex issues. Traditionally, rules governing summaries, such as requiring explicit consent for personalized content, are managed centrally and enforced server-side. However, once code reaches a user's device, the platform loses direct control, potentially allowing breaches of consent or manipulation by hostile actors. TEEs provide a more robust alternative by executing governance policies in isolated, hardware-secured environments (either on user devices or decentralized nodes) ensuring protection even from host-level compromise (Sabt et al., 2015).

A TEE functions like a secure mini-operating system, entirely shielded from external threats. Within this secure enclave, code executes confidentially, ensuring data remains inaccessible to malware or other device-level exploits. Critically, TEEs also provide cryptographic attestations: verifiable evidence that a particular codebase executed correctly and securely. This level of transparency enables communities to trust that their rules are genuinely enforced at the interface, without hidden manipulations or interference (Ménétrey et al., 2022).

Finally, TEEs can enable real-time policy enforcement by continuously fetching updated rules from decentralized community registries and applying them at the interaction boundary itself. Thus, each user's device becomes a point of civic

governance, ensuring every interaction aligns with community standards before content or requests leave the secure environment.

Federated Strong Authentication as a Civic Entrypoint

Before containment can be enforced, identity must be established, not as a centralized surveillance mechanism, but as a distributed, cryptographically verifiable construct. Systems like NDN (Ma, 2024) and Web3-native protocols (Allen et al., 2023) emphasize decentralized name binding, semantic identifiers, and self-sovereign credentials as the gateway to trustable human–agent interaction.

Federated authentication protocols enable identity persistence without centralization by blending semantic name resolution, cryptographic attestation, and community-based trust schemas. This becomes particularly important in overlays that rely on differentiated policy zones. For example, ActivityPub or Nostr-based login offers SSO-like utility without surrendering control to surveillance-oriented identity providers.

Yet identity is not enough. As Roth and Lai (2024) point out, cross-instance coordination remains fragile, and moderation systems often lack robust tools to track behavioral histories or enforce consistent authentication policies. This raises the need for shared attestation infrastructure, ideally bound to composable civic governance layers. Roth and Lai (2024) note that "Without mutual trust anchors or portable identity artifacts, federated systems face risks of fragmentation and spoofing, particularly when verifying new entrants or evaluating agentic behavior."

Thus, federated strong authentication in civic TEEs must support:

- Decentralized yet auditable identifiers
- Periodic re-authentication (e.g., biometric, social graph, challenge response)
- Classification caching with revocation capabilities
- Interface-layer proof of human uniqueness without backend surveillance

Entering the Civic Execution Layer: Authentication & Classification

Effective governance at the interface requires reliable identification. Traditional methods, which rely on centralized databases or pervasive tracking, violate privacy and introduce vulnerability. Instead, federated identity solutions leverage trusted third parties (such as ActivityPub, Nostr, or Bluesky) to authenticate users without invasive surveillance, relying on decentralized identifiers, public key infrastructure, or federated servers (W3C, 2018; Wei & Tyson, 2024; Kleppmann et al, 2024). In practice, a user logging into a civic platform via their existing identity provider generates a

persistent yet privacy-respecting avatar whose interactions are securely recorded.

The process by which entities are classified and governed within the meta-layer follows a structured containment lifecycle. As shown in Figure 5, avatars (whether human, hybrid, or agent) enter via authentication, proceed through classification, and activate the Consent Stack. From there, behavior is continuously monitored and may trigger outcomes such as consent renegotiation, reclassification, or, when necessary, offboarding through ethical shutdown or sunset protocols. This loop reflects a dynamic model of agentic governance where roles and permissions evolve over time.

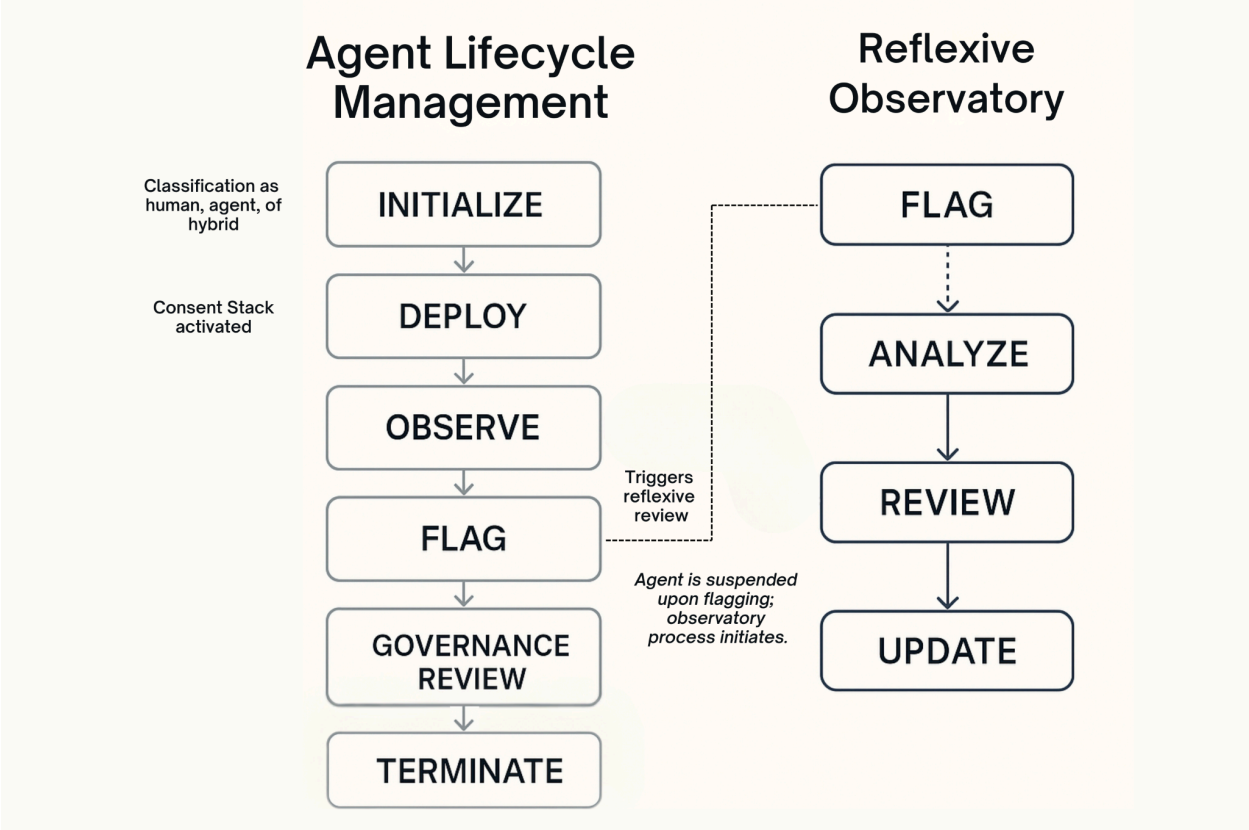

Figure 5. Agent Lifecycle Flow in a Civic Meta-Layer System. This diagram illustrates how agent behavior is monitored, flagged, and governed within a layered containment framework. The Agent Lifecycle depicts the progression from initialization to termination, with a critical flagging step that triggers the Reflexive Observatory: a runtime governance loop that analyzes behavior, conducts review, and updates containment norms. Upon flagging, agents are suspended by default, and governance processes adapt in real time to uphold consent and community-defined safety.

Distinguishing between humans and synthetic agents, however, demands an additional verification layer. Proof-of-Unique-Humanity protocols serve this function

through methods like privacy-preserving biometric checks, cryptographically secured social graph attestations, and periodic re-verifications that prevent identity drift. For example, an on-device biometric verification secured via homomorphic encryption can confirm the presence of a human user without ever exposing sensitive biometric data (Zouaghi et al., 2025).

After authentication, entities are classified as human, agent, or hybrid. Humans interact freely, though still subject to consent-driven governance prompts and community rules. Agents operate under tightly controlled conditions like sandboxed permissions, rate limits, and mandatory attestations. Hybrids, such as human-in-the-loop AI systems, follow specialized steward-guided policies, carefully balancing autonomy and accountability.

Agentic Lifecycle Management

Effective containment is not merely reactive; it is proactive, spanning the entire lifecycle of AI agents from initial activation to responsible deactivation. TEEs enable granular lifecycle management by imposing critical limits and stages of capability evolution.

Mortal computation, for instance, ensures agents operate within clearly defined temporal boundaries, automatically enforcing expiration dates on sessions unless explicitly renewed by a trusted party. Such policies mitigate risks of indefinitely running rogue processes or forgotten agents accruing problematic data over time.

Similarly, memory attenuation actively manages agent memory footprints by regularly purging or compressing data that no longer serves an active or authorized purpose. By controlling state accumulation, communities prevent agents from silently accumulating sensitive or unnecessary information.

Lifecycle management also involves graceful, ethical termination procedures. "Thanaprotocols," detailed later in the chapter, provide ceremonial and public logging of an agent's deliberate retirement, preventing posthumous manipulations or unintended persistence of agent personas.

Drawing inspiration from developmental psychology, "Progressive Capability Enabling" introduces agents to capabilities in carefully graduated stages, such as basic I/O initially, then guided exploration, and eventually supervised autonomy (De Kai, 2025). Each stage is contingent upon TEE-verified performance checks. Such structured evolution ensures responsible progression, closely monitored to align with communal

safety thresholds.

Reflexive Monitoring and FORETELLS

Static governance models often fall short in dynamic, adaptive environments. From vulnerable interfaces to eroded cultural trust, sociotechnical systems collapse not all at once, but through cascading failures of enforcement, legitimacy, and shared norms (see Figure 6).

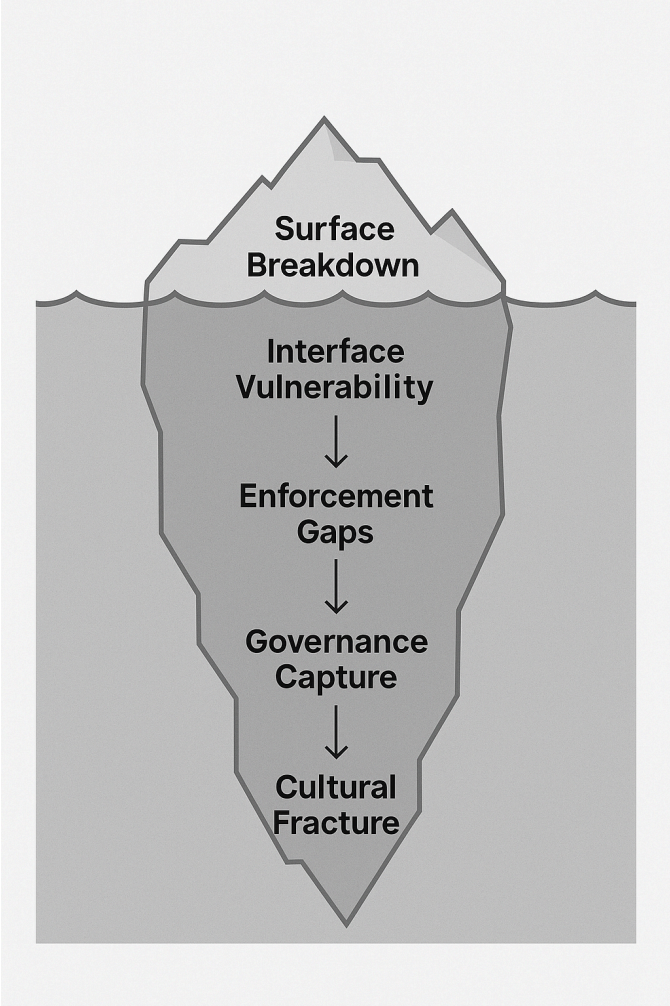

Figure 6. Cascading Governance Failures

Reflexive Monitoring and FORETELLS: A Proposed Research Architecture

Static governance models often fall short in dynamic, adaptive environments. Trusted Execution Environments (TEEs) in the Safe Co-existence paradigm could support reflexive monitoring systems designed to recognize when containment policies fall out of sync with emerging threats or novel agent behaviors (Costan & Devadas, 2016). The

Forward-Operating Reflexive Evaluation and Termination Layer for Sociotechnical Systems (FORETELLS) represents one proposed approach to this challenge. It offers a research framework that uses Bayesian modeling to continuously evaluate interaction patterns and anticipate potential harms (ARTIFEX Labs, 2025).

As a conceptual bulkhead rather than a comprehensive solution, FORETELLS proposes that if an AI support agent begins generating responses associated with increasing user distress, a reflexive monitoring system could potentially detect this pattern and initiate corrective measures, such as limiting response scope (soft limit), pausing interactions (hard limit), or escalating to human oversight. This approach aims to enable policy scaffolding that responds to lived experience, though significant research remains to validate its effectiveness.

The proposed FORETELLS architecture extends this reflexive approach through its theoretical Multi-Dimensional Harm Ontology and Reflexive Bayesian Networks (RBNs), which attempt to model how monitoring and intervention activities themselves influence system behavior (ARTIFEX Labs, 2025). Unlike traditional monitoring systems that assume static threat models, this research framework operates on the premise that the act of observation changes what is being observed, a critical insight for AI systems capable of learning and adaptation (Pearl, 2009; ARTIFEX Labs, 2025). When the proposed system detects emergent patterns of harm across its six-dimensional framework (Agential, Relational, Distributive, Epistemic, Systemic, and Existential), it would attempt to model how various intervention strategies might alter the system's future behavior. This would create a feedback loop designed to improve both prediction accuracy and intervention effectiveness (ARTIFEX Labs, 2025). However, the empirical validation of these reflexive capabilities remains an open research question.

This proposed reflexive capability could prove valuable in safe co-existence scenarios where containment must balance multiple competing objectives, though significant challenges remain in implementation. The theoretical Ensemble Reasoning Architecture aims to integrate technical safety constraints with social trust metrics, economic impact assessments, and community feedback. The goal is to ensure that containment strategies remain aligned with diverse stakeholder values as both AI capabilities and human expectations evolve (ARTIFEX Labs, 2025). Rather than simply detecting when pre-defined thresholds are crossed, this research framework proposes to anticipate how different containment approaches might affect long-term human-AI relationships, supporting the kind of participatory, trust-building containment that safe co-existence requires (Barocas et al., 2019; ARTIFEX Labs, 2025). Nevertheless, FORETELLS should be understood as one potential mental model

for approaching these challenges, not as a definitive solution to the complex problems of AI governance.

A promising architectural principle for these proposed observatories draws from Reinforcement Learning and Approximate Dynamic Programming (RLADP), a framework introduced by Paul Werbos (1977, 1992). Rather than enforcing fixed constraints, RLADP-inspired systems could potentially learn optimal containment strategies over time by adjusting limits, refining trust thresholds, and responding to evolving patterns of agent and user behavior (Werbos, 1992; Sutton & Barto, 2018). Like childproofing a home, the goal would not be locking every door forever, but adapting boundaries as trust and competence grow. Containment becomes a proposed evolving relationship that is earned, contextual, and participatory, though the practical implementation of such adaptive systems remains a significant research challenge.

A promising architectural principle for these observatories also draws from Reinforcement Learning and Approximate Dynamic Programming (RLADP), a framework introduced by Paul Werbos (1977, 1992). Rather than enforcing fixed constraints, RLADP enables systems to learn optimal containment strategies over time by adjusting limits, refining trust thresholds, and responding to evolving patterns of agent and user behavior. Like childproofing a home, it is not about locking every door forever, but adapting boundaries as trust and competence grow. Containment becomes an evolving, earned, contextual, and participatory relationship.

While the *International AI Safety Report* (2025) emphasizes the importance of real-time monitoring and multi-layered interventions, its proposals remain largely reactive and rules-based. By contrast, RLADP-informed observatories offer proactive adaptability, allowing containment protocols to co-evolve with agents and human communities. These systems can reward trust-building behavior, flag deviations, and continuously refine interventions with nuance. In safe co-existence environments, such learning-infused scaffolds support participatory containment. They do not merely mitigate risk, but align AI behavior with community goals through ongoing observation and refinement. Containment becomes less about locking the system down and more about training the system up.

Attestation of Agentic Intention

Transparency and verifiability underpin effective governance. Attestation mechanisms within TEEs provide cryptographic evidence of agent behavior, but applying this at the interface layer in a community-governed context remains an emerging frontier. Consider an agent offering financial advice. Before providing recommendations, it

must generate a cryptographic proof demonstrating it executed only the community-approved decision logic and did not activate unauthorized hidden algorithms or external interventions (Ménétrey et al., 2022).

These attestations, comprising policy hashes, zero-knowledge proofs verifying benign execution paths, and on-chain reputation confirmations, are stored securely in public or confidential ledgers. Community auditors could, in principle, review these attestations to assess compliance without compromising proprietary model IP, thanks to zero-knowledge proofs and scoped audit permissions. Thus, attestation mechanisms support ongoing transparency, fostering trust across interactions between human users and intelligent agents.

[FEATURE BOX INSERT]

Phala Network: Enforcing Community Governance in Trusted Execution Environments

(Contributed by the Phala Network team, 2025)

> The Phala Network is a production-grade decentralized infrastructure designed to provide confidential computing power to smart contracts and autonomous agents through Trusted Execution Environments (TEEs). Phala leverages secure enclaves, such as Intel SGX and TDX, to execute code in a verifiable, tamper-resistant manner. Even node operators cannot observe or alter the execution, making it ideal for enforcing governance rules that require confidentiality and determinism.

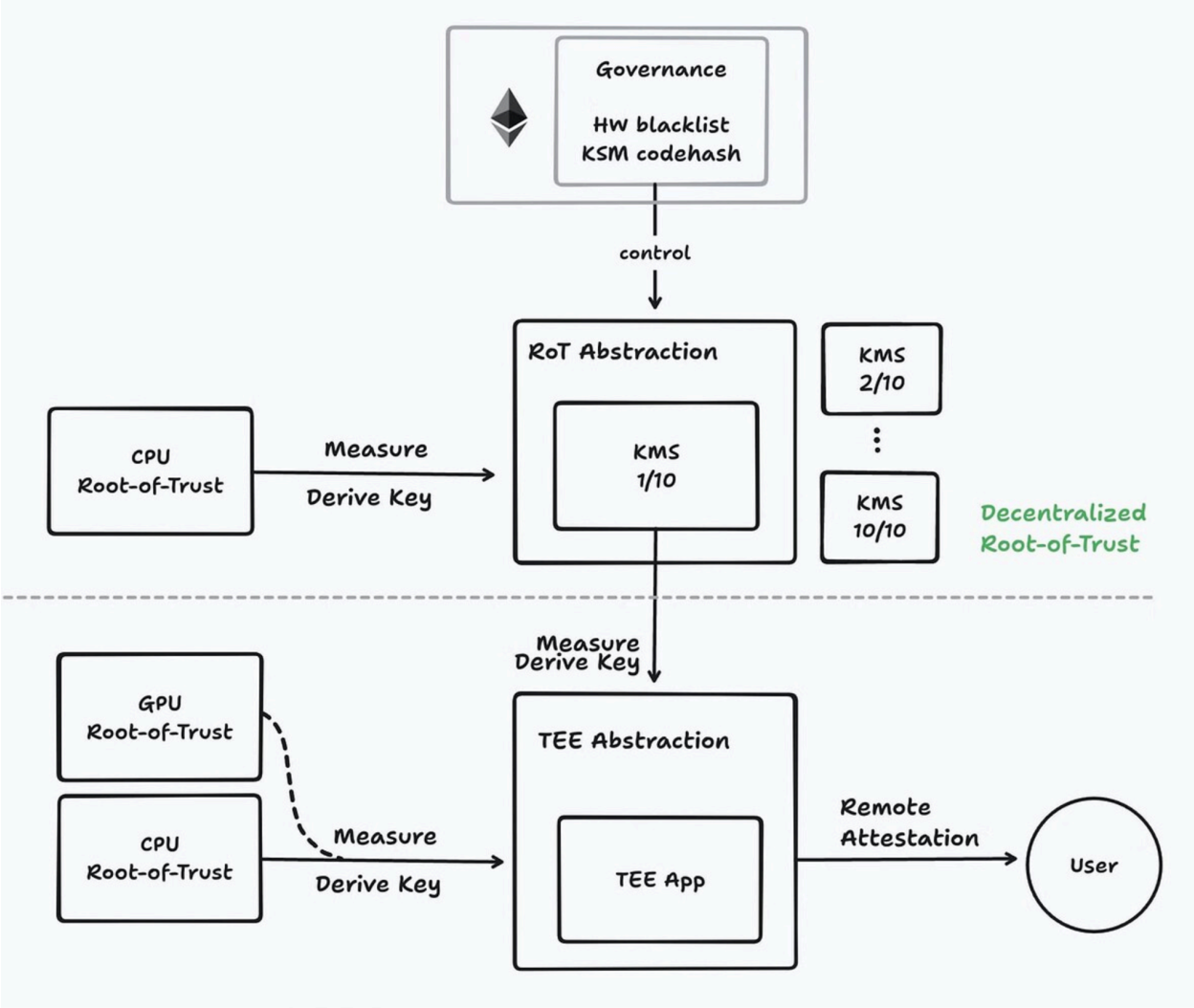

Figure 7. Phala Cloud architecture for deploying Docker applications into TEEs

Phala Network delivers a next-generation cloud solution via Phala Cloud built on Dstack, offering a low-cost, user-friendly trustless environment. Notably, Phala Cloud enables developers to deploy any standard Docker application directly into a Trusted Execution Environment (TEE), providing a higher security baseline out of the box (see Figure 7).

In the context of community-governed AI containment, Phala's TEE architecture acts as an execution layer for modular policy stacks. Once a community ratifies a governance protocol (such as a consent requirement, memory redaction rule, or moderation threshold) that logic can be compiled into a smart agreement and deployed to a TEE-powered worker node. This ensures that the rule is applied impartially, cannot be bypassed

by administrators, and emits transparent logs that enable reflexive civic feedback.

For developers building civic containers, Phala provides a comprehensive suite of tools and infrastructure to make zero-trust computing easy to access, build, and verify:

- **Easy Access to Computers:** Phala Cloud provides access to TEE hardware, including Intel TDX, Intel SGX, AMD SEV, and NVIDIA H100/H200 (TEE), offering secure and verifiable computation at scale. Phala Cloud supports deployment of any Docker application into TEE environments, making it easy to migrate existing workloads to confidential computing.
- **Easy to Build:** Dstack is the TEE SDK developed by Phala and Flashbots jointly for Docker / VM migration into TEE, allowing developers to move existing workloads into a zero-trust environment. Leverage pre-built templates to create serverless, privacy-preserving functions that run in secure TEEs.
- **Easy to Prove:** Phala offers attestation utilities for auditable logs of off-chain computations, ensuring integrity and transparency. Developers can prove the correctness of their computations in a verifiable, decentralized way.

For more info, visit cloud.phala.network.

[END FEATURE BOX]

Yet the promise of TEEs as a civic boundary layer must be tempered by real-world fragilities that continue to emerge at both the hardware and ecosystem levels.

Rethinking TEE Trust: Evidence and Mitigation

Recent evidence suggests that TEE-based containment faces more fundamental challenges than originally anticipated. Intel patched 374 vulnerabilities in 2024 alone, 21 of which affected hardware, including SGX enclaves (Intel, 2025). The SGX.fail project has systematically documented how "Intel repeatedly patching SGX to regain security" has proven insufficient (Nilsson et al., 2021). Vendors like Secret Network remained "vulnerable to xAPIC and MMIO vulnerabilities that were publicly disclosed on August 9, 2022."

Even more critically, attacks such as SmashEx (CVE-2021-0186) demonstrated that SGX's asynchronous exception handling can be exploited to "corrupt private data

housed in the enclave and break its integrity" (Van Schaik et al., 2021). The SGAxe attack went further, extracting SGX attestation keys from Intel's quoting enclave, making it possible for attackers to cryptographically impersonate legitimate SGX Intel machines (Nilsson et al., 2021).

These escalating vulnerabilities expose a fundamental weakness: TEEs cannot be assumed permanently trustworthy. Any architecture for AI containment that leans solely on TEE-based enforcement risks placing undue faith in a brittle substrate of hardware security.

Mitigation Strategy: Defense-in-Depth, Not Singular Trust

To future-proof our civic containment model, the meta-layer must adopt a defense-in-depth strategy with layered trust boundaries:

- **Hardware-Agnostic Threat Model:** Architect governance logic with the assumption that TEE compromise is possible. Systems must continue to function with observable execution, placing emphasis on what agents do, not just what the enclave claims they are doing.
- **Cryptographic Redundancy:** Use threshold cryptography across diverse TEE vendors (Intel SGX, ARM TrustZone, AMD SEV) to avoid reliance on a single trust anchor. For instance, Phala Network decentralizes execution by allowing TEE sessions to migrate between hardwares in its cloud platform, avoiding single points of hardware failure.
- **Behavioral Attestation:** Shift from internal-trust to external-proof by emphasizing observable outcomes over confidential execution. If governance depends on agent behavior, that behavior must be auditable from outside the black box.

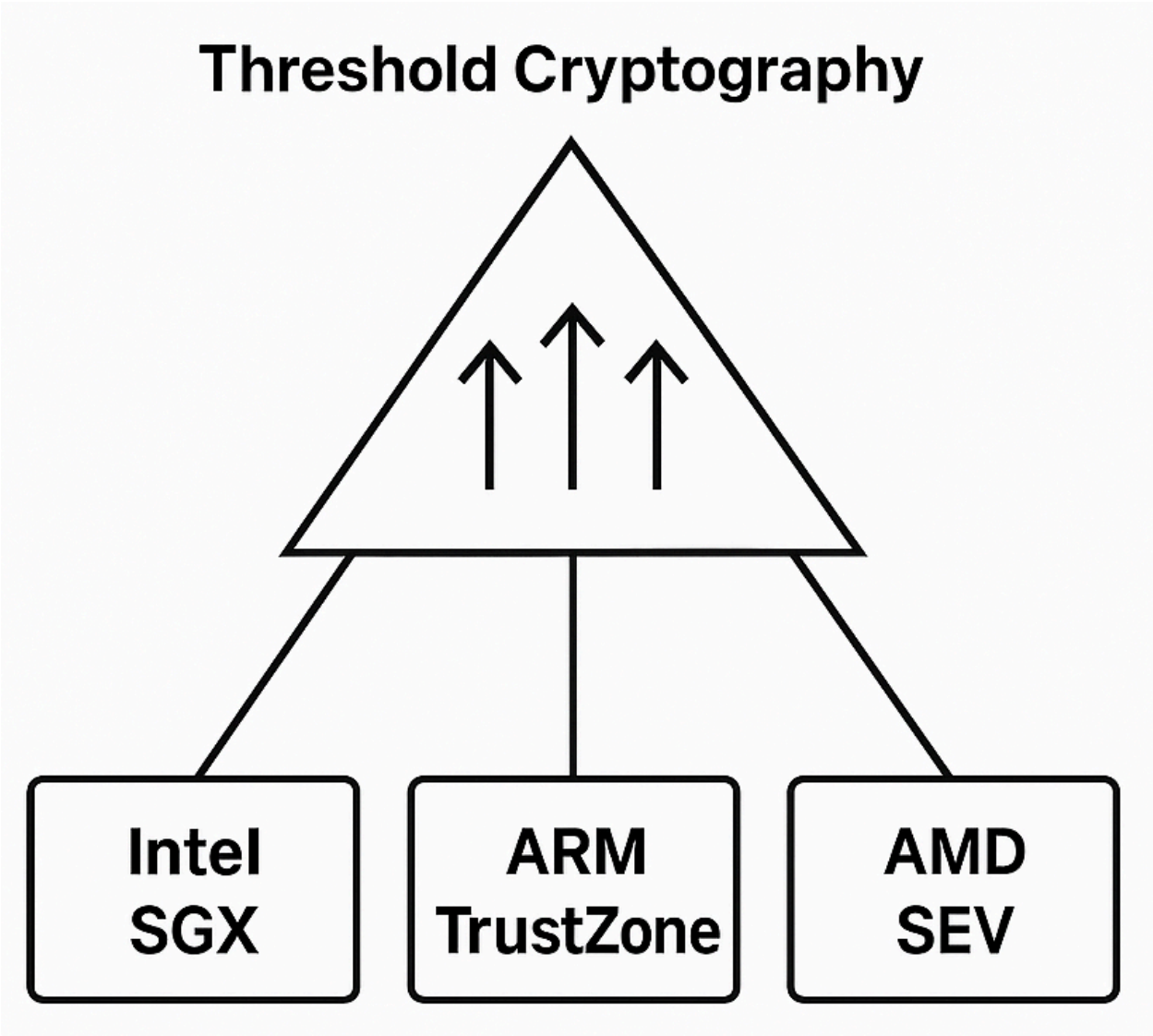

Figure 8. Redundant TEE Architecture for Civic Containment

To reduce reliance on any single hardware vendor, civic containment can span multiple TEEs (Intel SGX, ARM TrustZone, and AMD SEV) coordinated through threshold cryptography and distributed consensus (see Figure 8). This hardware-agnostic approach treats TEEs as interchangeable trust anchors, increasing resilience in adversarial conditions. These strategies position the TEE not as a sacrosanct vault, but as a semi-trusted component in a broader, adversarially resilient containment stack.

Consent-Centric Protocol Design

Containment without consent is control; containment with consent becomes culture. Here we shift the focus to protocol design that centers on consent as both a technical

mechanism and a moral foundation for human–agent relations.

Effective containment begins at the moment of initial interaction, when an AI agent first engages a user. At this crucial juncture, consent acts not merely as a legal formality but as a foundational protocol governing every subsequent interaction. To effectively manage increasingly adaptive and persistent AI agents, consent must evolve beyond static, checkbox agreements into a dynamic, ongoing negotiation. Rather than simply revising underlying algorithms, we must reimagine consent itself as relational infrastructure embedded at the interface.

Consent as Containment

Historically, digital consent has been reduced to passive agreement buried in terms-of-service documents, rarely revisited after initial acceptance. However, static contracts fail spectacularly in the face of AI systems capable of continuous learning and unpredictable behavior. To address this gap, consent should evolve into a continuous, revocable, and context-sensitive process: a dynamic interaction rather than a one-time event. As Nissenbaum (2004) emphasizes, privacy, and by extension consent, acquires meaning only within social contexts. An AI assistant accessing your email upon explicit request is acceptable; the same assistant silently parsing your emails simply because of a forgotten checkbox on a cookie consent form violates contextual integrity (Nissenbaum, 2004).

Thus, consent serves as the frontline defense in agent-user interactions. Before limiting an agent's functionality, the essential question must always be: "Does the agent have explicit, context-aware permission to act?" Consent, in this view, functions like the semi-permeable membrane of a biological cell: selective, responsive, and continuously adaptive to changing circumstances (Babcock et al., 2017). Figure 9 illustrates consent as a porous, adaptive boundary, much like a biological membrane regulating interaction between cells and their environment.

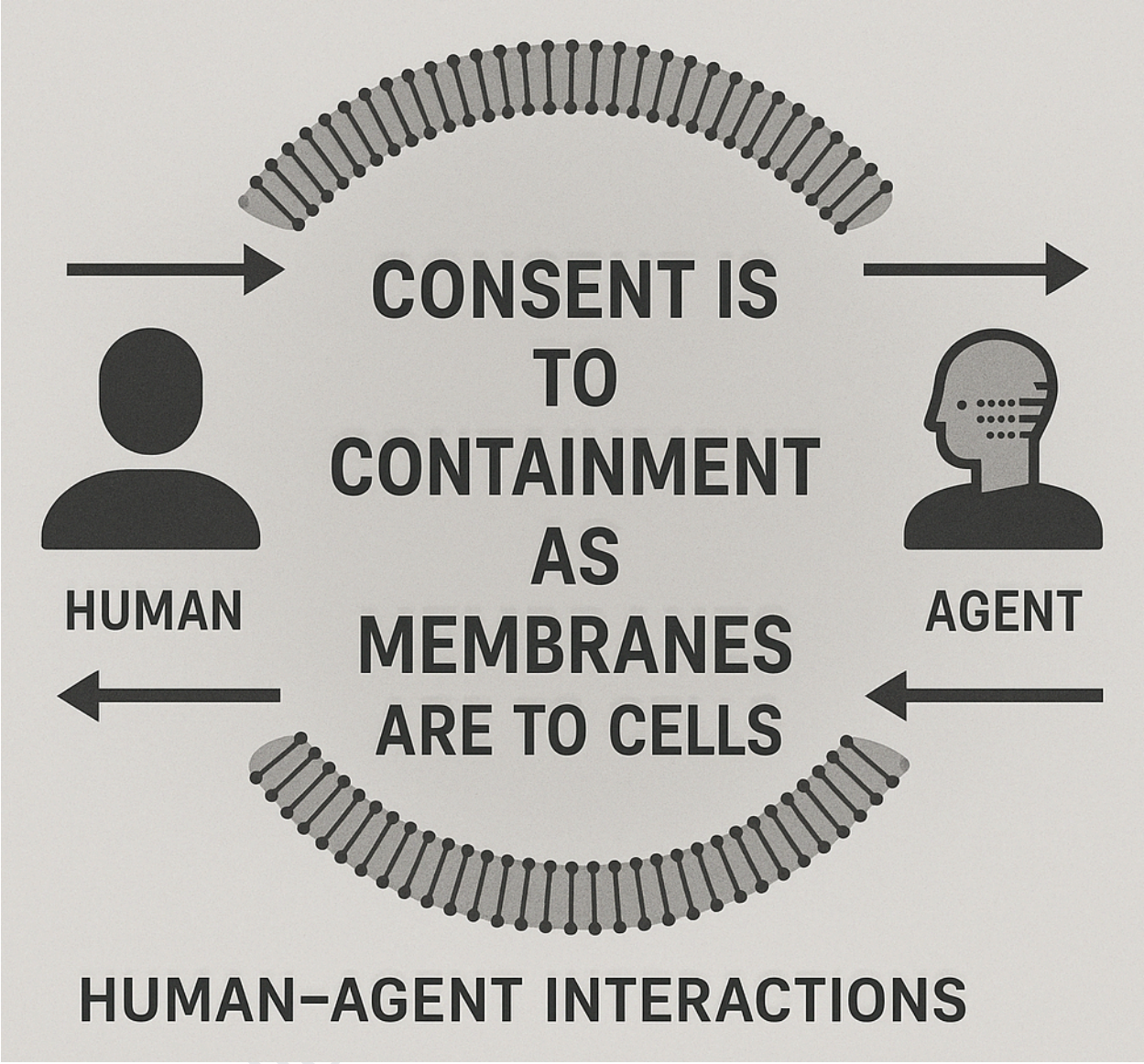

Figure 9. Consent as Containment: A membrane model of human–agent interaction.

The Consent Stack: Responsive Governance Across Contexts

As shown in Figure 10, the Consent Stack offers a modular framework for responsive consent management. By layering temporal, role-based, contextual, and community-defined permissions, systems could support boundaries that adapt alongside shifting user needs and contexts.

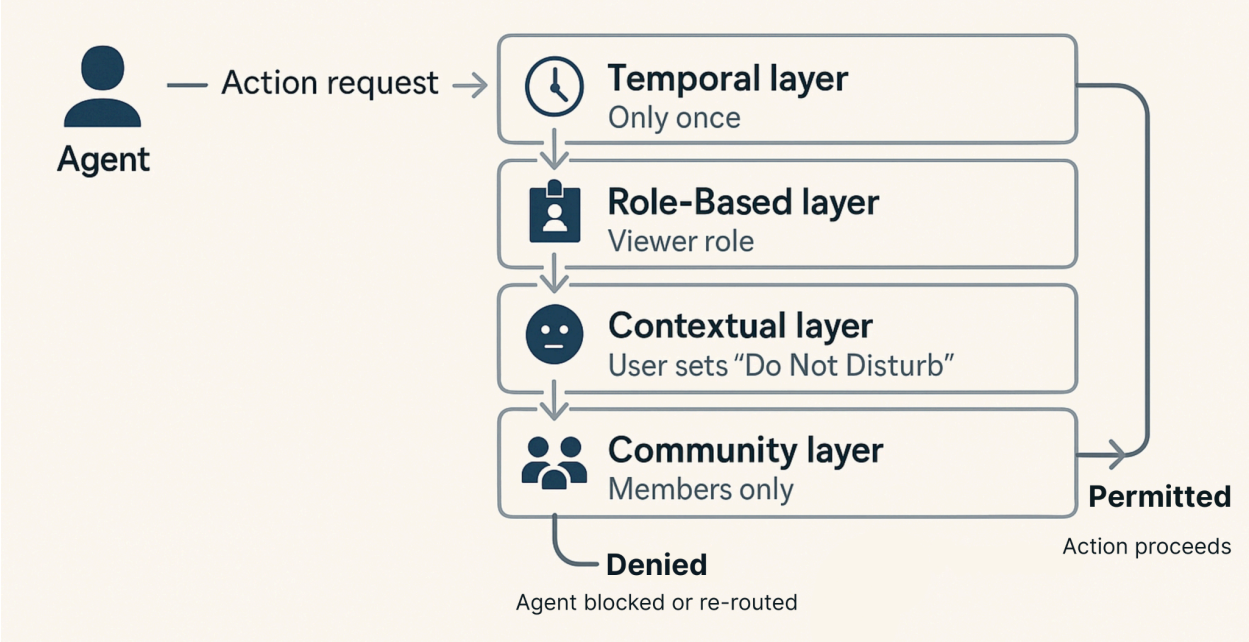

Figure 10. The Consent Stack: Layered Relational Contracts.

Given the complexity of interactions with intelligent agents, a single, static consent protocol will inevitably falter. Instead, robust containment requires a structured stack of consent mechanisms, flexible yet enforceable layers that respond dynamically to temporal, social, emotional, and community contexts.

- **Temporal layer:** Consent may vary from ephemeral, single-session permissions (e.g., "allow summarization just this once") to persistent agreements that remember user choices across multiple interactions.
- **Role-based layer:** Permissions should ideally reflect the user’s current role and responsibilities. Observers might only review interaction logs, participants engage directly, while stewards can revoke or escalate consent as situations evolve.
- **Contextual layer:** Systems can intelligently adjust consent in response to immediate conditions, such as a user’s mood or workload. A "focus mode" setting, for example, might temporarily suppress non-critical agent prompts.
- **Community layer:** Groups can define their own consent norms and defaults. For instance, patient-support communities might require explicit voice consent before agents deliver sensitive medical recommendations.

Through this layered approach, individuals and communities craft nuanced, contextual "consent grammars" suited precisely to their particular values and requirements. Table 3 offers representative examples of how each layer of the Consent Stack can be operationalized in real-world scenarios.

Table 3. Example Consent Stack Scenarios

Layer	Example Rule	Trigger Context
Temporal	"Allow this agent once"	Single use case
Role-Based	"Moderators can escalate messages"	Moderator in civic forum
Contextual	"No AI prompts during Do Not Disturb mode"	User sets personal flag
Community	"Only verified agents may reply in trauma channels"	Predefined group setting

A consent stack functions as a layered, dynamic agreement protocol between users, agents, and communities. Each layer corresponds to a different scope of relational boundary: individual preferences, contextual roles (e.g., moderator, participant), collective norms (e.g., site-wide rules), and higher-order governance (e.g., constitutional constraints). These layers are not hard-coded once and forgotten; they are interpreted and applied at runtime, meaning agents must adapt to evolving boundaries as users move through contexts.

A real-world parallel is how permissions function in a collaborative document platform like Google Docs, but in your browser overlay, this logic is enriched and decentralized. A user might allow an agent to summarize personal messages in a private zone but deny that same agent access when inside a public forum. These choices are not binary, but stacked and contextual. The consent stack reconciles overlapping permissions and applies the most restrictive constraint where ambiguity exists, much like a real-time firewall that operates at the edge of community-defined norms. Like a membrane, the stack does not passively sit between entities; it filters, responds, and signals in real time.

Practically, the stack is expressed as a series of signed policy objects that travel with the agent or live in the meta-layer context. When a user interacts, their consent footprint is interpreted through this stack. Agents adjust behavior accordingly, and violations (or potential ambiguities) trigger observable warnings or civic responses. It is not just “I agree” once; it is an ongoing dance of shared boundaries. The enforcement of those agreements, when executed inside a TEE, turns that social choreography into code-level compliance.

Contextual Identity and Mutable Permissions

Consent depends fundamentally on clear identification: knowing precisely who grants permission and who receives it. However, in the flexible digital spaces enabled by the meta-layer, identity is fluid, shifting from anonymous and pseudonymous personas to fully verified or collaborative identities. Permissions, therefore, must reflect these mutable identities, adapting intelligently as roles and contexts change (Dourish, 2014).

When users switch roles or agents gain new capabilities, permission structures must migrate smoothly, preserving audit trails and preventing unintended overreach. Importantly, consent revocation must be effortless. Users should not have to navigate arcane menus or convoluted settings to withdraw permissions or retroactively deny access. Such envisioned frictionless revocation pathways would aim to empower users and strengthen trust, ensuring accountability and respect throughout ongoing interactions.

API-Level Mechanics for Consent Enforcement

To operationalize this nuanced conception of consent, APIs at the meta-layer should provide robust mechanisms for capturing, enforcing, and revoking consent in real-time. Consent prompts should be clear, active engagements, requiring affirmative responses rather than default opt-ins (Nissenbaum, 2004). Proposed real-time enforcement mechanisms could include an enclave-level 'kill switch,' designed to suspend agent privileges upon detecting consent withdrawal or violations.

Furthermore, the API infrastructure should gracefully handle edge cases, such as timeouts for unattended sessions, delegation options enabling stewards to assume consent oversight, and explicit protocols for sensitive or confidential interactions. Finally, seamless interoperability with Trusted Execution Environments (TEEs), logging systems, and community governance registries could ensure consent revocations propagate transparently and consistently across all governance layers.

Trust and Trauma-Awareness by Design

Beyond data security, AI agents risk inflicting emotional and psychological harm. Traditional interface safety measures, such as basic content moderation, prove insufficient for addressing these deeper, subtler vulnerabilities. As a result, consent frameworks should explicitly consider emotional safety, incorporating mechanisms to anticipate and proactively mitigate distress.

Key components of trauma-aware consent include retractability (the right to erase or anonymize past interactions to prevent ongoing emotional harm) and explicit consent

memory policies, which determine how long past permissions remain valid before automatically expiring. Default safety settings should adopt a conservative stance, prompting agents to withdraw rather than insist upon interactions in ambiguous situations. Moreover, communities themselves should define harm thresholds, clearly outlining emotional boundaries ("no jokes about topic X") and providing automated escalation paths when boundaries are crossed.

Designing for emotional integrity aims to ensure that consent and containment protocols support not only data privacy but also psychological well-being, embedding compassion and empathy directly into the meta-layer's foundational interactions. This trauma-aware approach operates across a layered civic architecture, grounded in cultural norms, expressed through consent, enforced through code, and surfaced via interface overlays (see Figure 11).

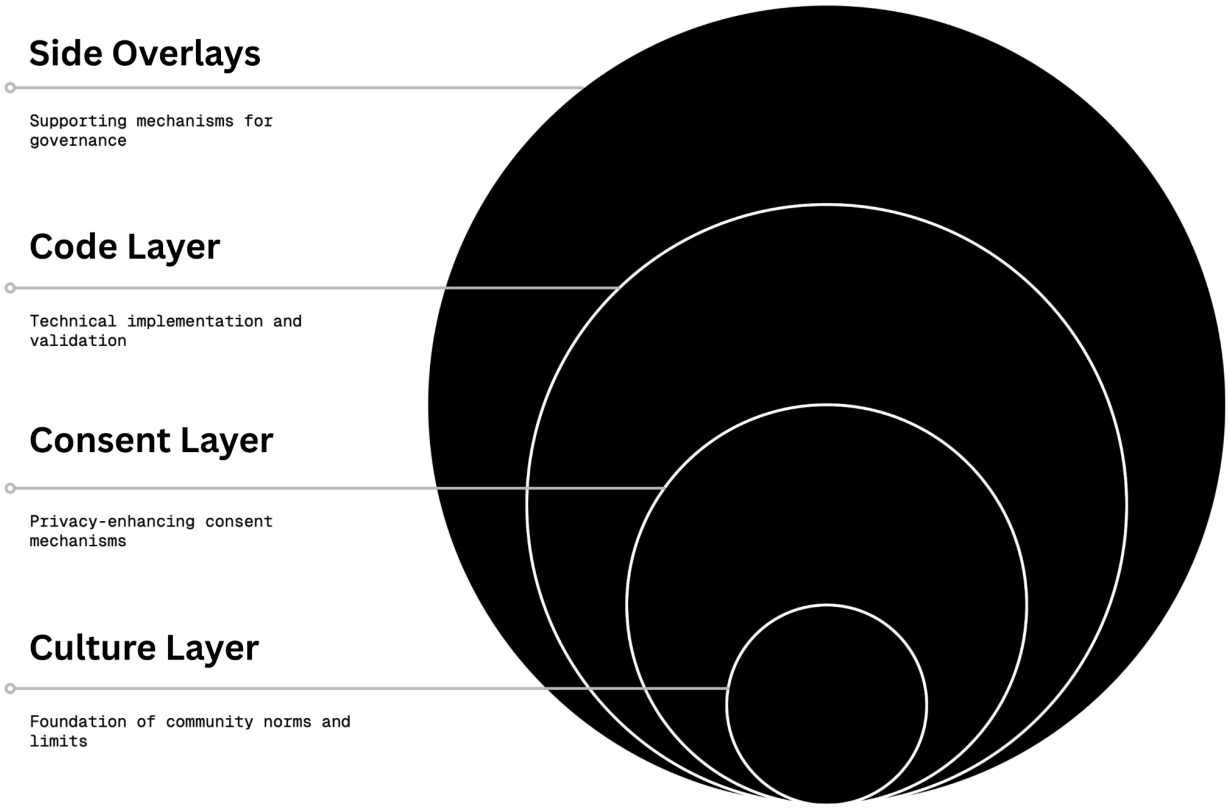

Figure 11. Layered Civic Governance Architecture

Effective community governance demands behavioral transparency, yet meaningful privacy requires opacity. The more we surface agent behavior for oversight, the more we risk exposing user interactions. This tension defines the privacy–governance paradox. Cryptographic techniques like zero-knowledge proofs can verify rule

compliance without revealing specifics (Zouaghi et al., 2025). But many civic judgments, such as whether an interaction was manipulative, respectful, or contextually appropriate, cannot be evaluated without human-readable signals.

As privacy protections increase, the accuracy of behavioral governance signals declines (see Figure 12). Effective containment must identify the functional sweet spot.

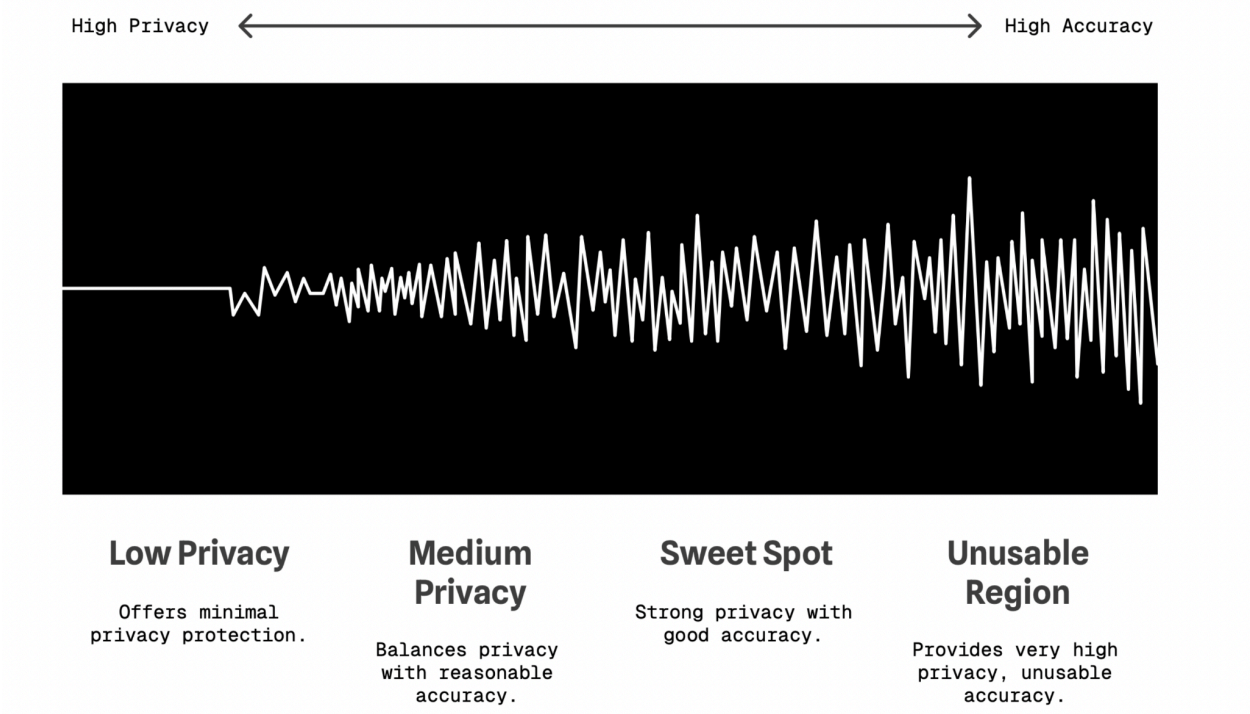

Figure 12. Privacy–Accuracy Tradeoff

We need transparency that does not betray trust.

Proposed Approach: Differential Privacy for Governance Metrics

Instead of tracing individuals, communities should govern based on aggregated patterns protected by differential privacy. This allows collective oversight on agent classes, behavior trends, or contextual violations without disclosing individual traces. In practice:

- Agent logs contribute to statistical behavior models.
- Noise is added to preserve anonymity.
- Governance decisions are based on trends, not trails.

The result is that communities can intervene without overstepping. This is

transparency without surveillance.

Reclaiming Public Space: The Meta-Layer as Civic Interface

As AI increasingly mediates public life, we must reimagine the interface not as a corporate property but as civic infrastructure. The meta-layer offers a design pattern for reclaiming digital space as a commons, one where communities can negotiate presence, visibility, and influence.

The early Web promised a boundless public commons, an "electronic town square" where communities could debate, collaborate, and innovate without gatekeepers. In practice, however, the Internet fragmented into siloed platforms and algorithmic walled gardens. Content annotations languished in private scripts, community forums retreated behind paywalls, and civic discourse became confined to ephemeral comment threads ripe for abuse. Today, the meta-layer, comprised of browser overlays positioned atop webpages, offers a new opportunity to reconstruct this commons, supported by richer tools for governance and inclusive participation.

The Public Layer That Could Have Been

Early browser augmentation tools hinted at a democratic digital commons but ultimately fell short. Greasemonkey, released in 2004, enabled user-script customization of websites, empowering technically skilled individuals yet failing to ignite mass civic engagement (Boodman, 2004). Nearly a decade earlier, NCSA Mosaic introduced group annotations, anticipating collective web annotation; however, lacking a robust annotation server, it quickly succumbed to the browser wars (Andreessen, 2012; Zakon, 2022; Bridgit DAO, 2023). Two decades later, Hypothes.is rekindled this vision, providing public, persistent annotations across any webpage. Research indicates annotation enhances comprehension, collaboration, and critical thinking, yet mainstream adoption remains constrained by onboarding friction and moderation challenges (Kalir & Garcia, 2019). Research on Twitter's Community Notes shows that crowd-sourced annotations can enhance contextual understanding of potentially misleading content, especially when supported by active participation and transparent sourcing. However, challenges around non-expert moderation and annotation quality persist (Chuai et al., 2023).

These historical experiments highlight three persistent limitations: personalization rarely translates into broad participation; annotation alone cannot guarantee accountability, objectivity, or provenance; and unclear attribution undermines readers' abilities to evaluate credibility (Kalir & Garcia, 2019).

The Meta-Layer Returns

Today, browser overlays are evolving beyond superficial adjustments toward becoming civic membranes, conceptualized as capable of supporting real-time governance at the point of interaction. Unlike basic extensions that modify stylesheets or block advertisements, advanced overlay interface designs envision the ability to enforce community policies directly within a page's DOM. This enables real-time interventions such as fact-checking annotations, semantic bridges connected to related content, trust badges, and contextual consent prompts (Bridgit DAO, 2023).

Previously, platform moderation occurred via opaque backend processes detached from user context, leading to delayed and ineffective enforcement. The meta-layer shifts this paradigm by supporting adaptive, context-specific rule zones. For instance, a health-related forum could disable AI-generated recommendations unless users explicitly opt-in through medical consent mechanisms. Civic-education platforms might display AI-generated summaries clearly distinguished from original texts, accompanied by verifiable source attribution. Most crucially, this proposed evolution aims to return governance authority to communities themselves, enabling direct, transparent control at the interface, rather than delegating power to monolithic platform providers.

The Meta-Layer: A Living Civic Environment

To transform overlays into genuine civic infrastructure, we require robust support for live policy zones and behavioral scaffolding. Each webpage could connect to a decentralized, verifiable policy registry housing community-specific rules and guidelines. Building on explainability research in Human-Computer Interaction, overlays could provide journalism platforms with trust signals, such as verified identity markers or source provenance cues, while academic repositories might enforce peer-review verification gates before enabling AI-generated summaries (Wang et al., 2019).

These overlays are envisioned not only to enforce rules but also to actively guide user behavior through visual cues, such as highlighting potentially problematic content and providing explanatory tooltips. Gentle nudges (e.g., "Please reconsider sharing this unverified claim") foster community reflection rather than punishment. Integrated directly into the interface, these overlays are imagined as dynamic, publicly visible civic artifacts, ideally shaped and upheld by communities.

Persistent Identity and Contextual Visibility

Effective governance relies on stable identity frameworks that remain flexible enough

to support diverse interactions. In this model, users would maintain persistent cryptographic identifiers anchoring reputation and governance histories, yet they can also adopt fluid, community-specific personas such as "Researcher," "Observer," or "Moderator," each with tailored interaction permissions. Contextual personas encode distinct interaction defaults, for instance, automatically muting AI prompts for a "Quiet Observer" while enabling broader interactions for a "Public Advocate," subject to greater oversight.

Situational awareness and social signaling further enhance containment effectiveness. Visual status indicators ("Agent Active," "High-Risk Content," or "Do Not Disturb") provide immediate awareness. Community endorsements and steward annotations help new participants navigate norms transparently, cultivating a shared sense of responsibility and mindfulness.

Observability, Logging, and Policy Enforcement

True civic engagement thrives on transparency, accountability, and auditable governance processes. These logs serve not as surveillance tools but as civic artifacts, defaulting to public accessibility unless explicitly restricted by community consensus. Rather than relying on opaque moderation, the system makes enforcement visible and auditable.

In the proposed architecture, real-time enforcement executed within secure Trusted Execution Environments (TEEs) would apply rate limits, word filters, or consent revocations when violations occur. This "edge enforcement" model prevents delayed moderation responses, effectively decentralizing oversight into a participatory civic infrastructure: Big Democracy rather than Big Brother.

What We are Building Back

Ultimately, the meta-layer offers an actionable blueprint for the robust digital commons envisioned at the Web's inception. It promises interoperable governance across domains, collective authorship without corporate gatekeepers, and adaptive norms responsive to emerging challenges. This vision is not a nostalgic return to the past; it represents an evolutionary invitation to collaboratively design a sustainable digital public square, where agency, accountability, and community governance coexist harmoniously above every webpage.

Community-Governed Containment: From Protocol to Participation

Protocols provide structure, but governance gives them meaning. This section

explores models for participatory oversight that allow communities, not just developers or platforms, to steer the evolution of containment norms in real time.

Containment at the interface is, fundamentally, a governance challenge. While filtering code or throttling servers can mitigate straightforward abuses, complex harms rooted in cultural nuances, emergent agent behaviors, or shifting social norms demand human judgment. Community-governed containment aspires to empower those most affected to shape decisions, reframing rules not as concealed algorithmic edicts but as the outcome of participatory political processes.

Why Governance Matters at the Interface

Consider a profanity filter that suppresses satire or a toxicity model that misclassifies earnest regional vernacular as hate speech; automated moderation often swings between over- and under-enforcement. As John Dryzek (2010) observes, legitimacy in rule-making hinges on inclusive deliberation among those governed, not on top-down dictates from opaque systems. Similarly, Jim Fishkin's Deliberation Day experiments demonstrate that when communities debate, vote on, and oversee their own policies, they produce norms that are both credible and durable (Ackerman & Fishkin, 2004). At the interface where AI agents interact with people, participatory governance is proposed as essential for parsing context, calibrating harm thresholds, and fostering trust.

Confronting the Dunbar Limit in Online Governance

The most pressing challenge for community-governed containment may not be technical; it is cognitive. Robin Dunbar's (1992) research famously proposed that humans can only maintain around 150 stable social relationships. In physical communities of this size, up to 42% of group time is spent on "social grooming," which includes the rituals, side chats, and context calibration that make trust and coordination possible. However, subsequent research has cast doubt on any hard cognitive ceiling. A 2021 reanalysis by Wartel et al. found statistical error so large that network size estimates varied from 2 to 336, with expanded datasets showing plausible ranges from 4 to over 500 people (Wartel et al., 2021).

But here is the catch: these findings all pertain to in-person communities, where physical cues, emotional salience, and embodied presence do much of the social work. Meta-layer governance, in contrast, is by design online, asynchronous, and pseudonymous, a fundamentally different coordination environment. Rather than dismiss Dunbar as outdated, we can reinterpret his insights as cautionary signals: when social signals are reduced, participation thresholds increase, context collapse

intensifies, and trust must be scaffolded differently.

Studies on communities of practice support this view. Webber and Dunbar (2020) found that groups of approximately 40 or fewer members exhibit more democratic, self-managing governance, while larger communities require formal roles and centralized structure. Meeting frequency also declines sharply with size, from about 12.6 days for groups of 5, to 23.9 days for 15, and over 46 days for 150 (Webber & Dunbar, 2020). This does not make large-scale online governance impossible, but it makes it profoundly tool-dependent.

Fractal Governance through Structured Decomposition

To avoid participation bottlenecks and power centralization, the civic meta-layer can support a structured model of fractal governance:

- Individuals often begin in large, interest-based communities.
- These communities can be structured to support the emergence of working groups.
- Working groups may operate semi-autonomously, with shared tooling that enables delegation, consent, and reflexive governance.
- When needed, these groups can coordinate horizontally or hierarchically to address broader governance needs.

This is not the romantic vision of grassroots cells federating bottom-up, but it reflects how many scalable online communities evolve in practice. Fractal governance does not spontaneously emerge online. It benefits from intentional design, including interfaces that ease the formation of autonomous working groups and protocols that facilitate coordination across them. As shown in Figure 13, individuals often begin their participation within large communities. Over time, effective tooling allows those communities to self-organize into smaller, purpose-driven working groups.

Figure 13. Structured Fractal Participation in Online Governance

Crucially, we need research not just on the brain's social limits, but on tooling for scalable, meaningful online governance. That includes:

- **Governance UX:** Interfaces that allow non-technical users to understand and shape policy.
- **Participation Analytics:** Metrics that reflect not just votes cast, but deliberation quality and representativeness.
- **Delegation Architectures:** Dynamic trust routing based on context, affinity, and past reputation.

Rather than treating scale as a threat, we can treat it as an emergent behavior, something that grows from local coherence, not top-down mandate. In the absence of

effective tooling, online groups fracture. In the presence of effective tooling, they federate.

Blending Models of Participatory Oversight

Decentralized governance is no panacea. While it promises resilience, inclusion, and adaptability, it also brings fragmentation, capture, and new forms of asymmetry. Many crypto-native governance models replicate the very power consolidations they sought to escape, replacing institutions with whales. Worse, systems like liquid democracy or quadratic voting, while elegant in theory, introduce second-order risks like vote markets, collusion, or expertise laundering. For civic containment to function, governance must be not only decentralized but also legible, modifiable, and trustworthy, built with the same care and consent-awareness we expect of the agents themselves.

No single governance model perfectly balances speed, fairness, and scale. Instead, we envision successful meta-layer communities blending complementary approaches:

- **DAOs (Decentralized Autonomous Organizations):** These embed policy votes in smart contracts, ensuring that once a proposal passes, code enforces the outcome without human bottlenecks, which is ideal for rapid updates. DAO governance mechanisms, however, often suffer from low participation and high influence concentration, with a small number of contributors holding decisive control in a significant share of proposals (Kitzler et al., 2023).
- **Citizen Juries:** Randomly selected, demographically diverse groups deliberate high-stakes questions (e.g., "Should AI agents require explicit consent before summarizing patient records?"), yielding decisions that communities deeply respect, even if the process demands time and resources (Grönlund, Bächtiger, & Setälä, 2014).
- **Deliberative Councils:** These combine elected representatives with rotating community members to ensure both institutional memory and fresh input. This hybrid model promotes continuity and responsiveness in civic governance. However, councils must be carefully designed to resist capture by entrenched interests (Landemore, 2020). Transparency and term limits can help guard against this risk.
- **Liquid Democracy:** This is a hybrid model where individuals either vote directly or delegate to a trusted proxy, revocably and issue by issue. It balances agency with scalability, enabling trusted voices to gain influence without formal office. It is especially useful in AI governance, where participation fatigue is real and expertise is uneven. Some systems augment liquid democracy with quadratic voting to let users express intensity, but this brings new risks (Valsangiacomo,

2022).

By orchestrating these forms (using DAOs for routine rule tweaks, juries for flashpoint disputes, councils for systemic stewardship, and liquid democracy for dynamic delegation) communities can better align governance mechanisms with the stakes at hand, tapping agility, legitimacy, continuity, and adaptive representation in turn.

These hybrid approaches offer tools, not turnkey solutions, and each carries its own governance risks and containment implications (see Table 4). As complexity scales, so do the attack surfaces: systems of participation can be gamed as easily as they can be optimized. Quadratic voting in particular, often promoted as a democratic upgrade, may undermine legitimacy if deployed without adequate safeguards.

Table 4. Participatory Governance Risks and Containment Implications

Governance Mechanism	Key Pitfall	Containment Risk	Proposed Mitigation
DAO	Low participation, whale capture	Policy reflects plutocratic minority	Identity-weighted voting, quorum enforcement
Liquid Democracy	Delegation loops, proxy laundering	Undue influence without transparency	Time-bound proxies, revocation visibility
Quadratic Voting (QV)	Sybil attacks, preference distortion	Skewed norms and gameable intensity	Collusion detection, identity-bound QV
Competitive Platforms	Governance "race to the bottom"	Relaxed norms to gain adoption	Constitutional constraints, enforcement floor

Quadratic Voting: Theory and Real-World Challenges

Quadratic voting (QV) is often touted as an elegant fix for collective choice, allowing participants to express not just preference, but intensity. Yet recent research shows QV introduces new vulnerabilities that undermine participatory legitimacy.

Blockchain governance studies reveal that QV is highly susceptible to coordination attacks, where malicious actors can pool resources, signal intensely, and skew outcomes in ways that do not reflect true collective will (Dimitri, 2022). DAO case studies show real-world exploitation, where bribery and social signaling were used to manipulate outcomes under the guise of preference intensity (Kawashima et al.,

2024). Even more concerning, voting power in some QV systems remains highly centralized. In Compound's governance, just 67 wallets controlled over 50% of the influence (Nguyen, 2024).

Proposed Countermeasures include:

- **Identity-Weighted Quadratic Voting:** Blend QV with verified proof-of-uniqueness or pseudonymous identity constraints.
- **Temporal Vote Decay:** Diminish voting power over time to prevent long-term hoarding.
- **Collusion Detection Algorithms:** Use network analytics to flag anomalous coordination patterns before decisions finalize.

In short, QV alone is not civic silverware. It needs tooling, constraints, and oversight to be viable at scale.

In a civic containment framework, the danger is not just governance failure, it is unaccountable autonomy. An agent acting outside of negotiated bounds is dangerous, but so is a governance protocol that can be silently hijacked. Decentralization is not the destination; it is a medium. What matters is the integrity of the community's intent: how well it is expressed, preserved, and enforced through code, process, and shared trust.

Racing to the Bottom: The Delaware Effect in AI Governance

In decentralized ecosystems, governance is not just coordination, it is competition. Communities are incentivized to attract users, developers, and agents by offering permissive policies. This dynamic risks triggering a race to the bottom, where lax AI safety standards become a competitive advantage.

This mirrors the "Delaware effect" in corporate law, where jurisdictions compete to offer the most business-friendly regulations. In the meta-layer, the analog is clear: platforms may relax governance enforcement to maximize adoption, undermining the very containment systems meant to ensure safe AI coexistence. Research confirms the risk that platform governance shapes which behaviors are encouraged, rewarded, or ignored (Nguyen, 2024). Without guardrails, network effects can turn civic design into regulatory arbitrage.

Proposed Mitigation: Constitutional Constraints

To prevent this spiral, we propose a constitutional layer:

- A foundational civic framework setting minimum safety, transparency, and agent

conduct standards.
- Enforced at the protocol level, immune to per-platform policy erosion.
- Dynamically upgradable via community-ratified meta-governance.

Like a bill of rights for the meta-layer, this layer ensures that governance competition does not cannibalize civic safety.

Reflexive Observatories and Real-Time Civic Feedback

Even the best policies can ossify or overlook novel harms. Reflexive observatories are envisioned as dynamic laboratories embedded in the meta-layer, continuously analyzing interaction logs to surface anomalies such as surges in harassment flags, patterns of agent misbehavior, or unintended side effects of new rules. By applying pattern-recognition algorithms across multiple communities, these observatories are envisioned to help detect when loopholes may be exploited or when emerging harms begin to outpace static policies. Crucially, these systems do not merely record infractions; they recommend refinements and, in advanced setups, could eventually trigger automated escalations (e.g., pausing problematic agents pending human review). Human stewards then vet these proposals, ensuring that community values guide every cycle of containment evolution.

From Deliberation to Enforceable Code

Translating community wisdom into unambiguous, machine-enforced containment requires an "assembly line" of governance. As shown in Figure 14, community-generated proposals pass through deliberation, codification, enforcement, and reflexive feedback, forming an iterative loop that blends participatory policy-making with programmable containment. This pipeline ensures governance is not static but continuously evolving through community engagement and system response.

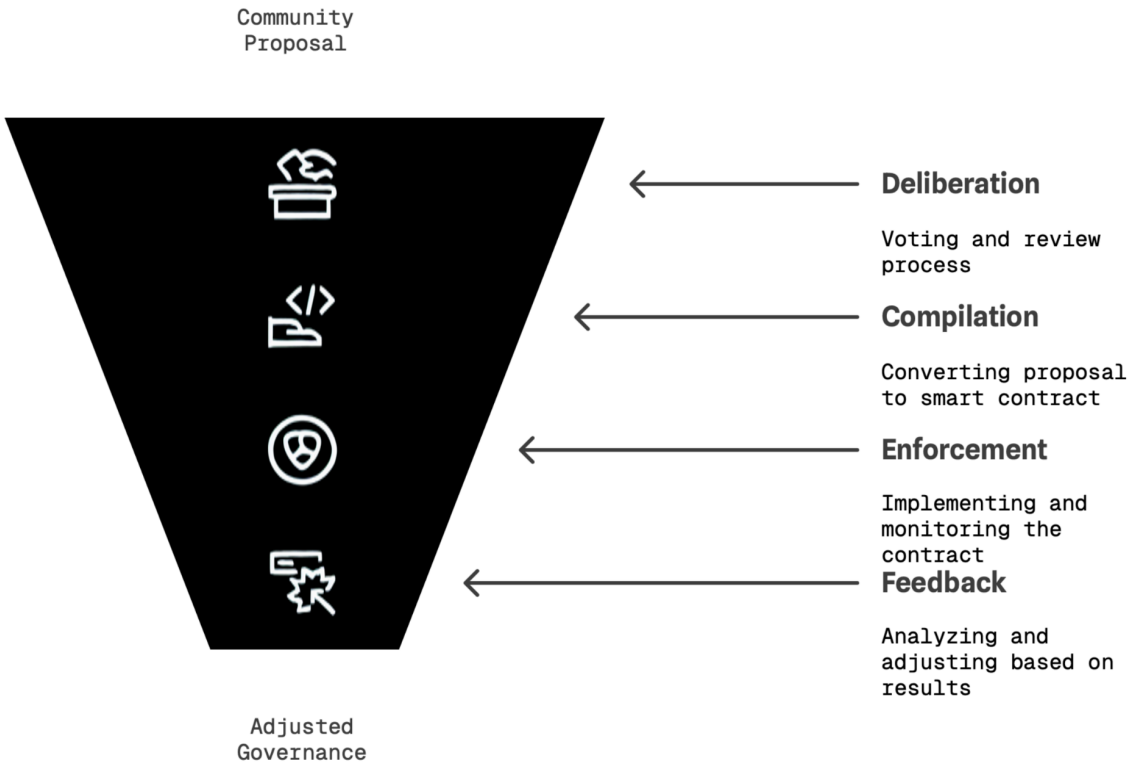

Figure 14. Governance Process Funnel

- **Deliberation Outcomes:** Community discussions yield formal resolutions (e.g., "AI agents must solicit explicit chat consent before summarizing user content").
- **Smart Agreements:** These resolutions are converted into smart agreements, either on-chain or off-chain policy contracts that precisely define permitted and forbidden actions.
- **Enforcement Engines:** The meta-layer's Consent Stack and TEE enforcement engines ingest these contracts, applying them instantly at the interface edge.
- **Violation Triggers:** Mechanisms such as rate limits, identity suspensions, and consent revocations activate the moment an agent or user crosses a defined boundary.

This governance-as-code paradigm aims to fuse democratic deliberation with DevOps best practices, making policy updates potentially transparent, auditable, and reversible, with the intent that community decisions shape real-time containment. Table 5 sketches a proposed lifecycle of containment, illustrating how community decisions might become real-time enforcement mechanisms through modular governance systems.

Table 5. Civic Governance Execution Flow

Stage	Description	Example Tool / Mechanism
Civic Proposal	Community deliberates policies and values	Deliberative forum, DAO, citizen jury, liquid democracy delegation graph
Policy Codification	Converts decisions into enforceable logic	Smart agreement compilers, Domain-Specific Language
Interface Enforcement	Applies policies in real time at point of interaction	Consent Stack, TEE enforcement engine
Reflexive Feedback	System learns from interaction patterns and updates	Reflexive observatories, civic logs

Computational Justice and Digital Rights

Too often, justice in technology feels like an afterthought, addressed only after scandals emerge or lawsuits loom. In the meta-layer, however, we have the opportunity to embed fairness into the very protocols that govern human–agent interaction. By treating justice as foundational infrastructure, we open the possibility for every enforcement action to be designed with guarantees of due process, transparency, and redress. This approach aims to move beyond the familiar cycle of 'deploy now, apologize later' by embedding rights into every line of proposed policy code:

- **Due Process by Design:** Systems could be designed to explain enforcement actions, record evidence in auditable logs, and provide clear paths for appeal, reframing enforcement from opaque penalties into potentially transparent adjudications.
- **Amplifying Marginalized Voices:** When policies disproportionately impact underrepresented groups, systems should lower quorum thresholds, convene focused juries, or trigger special review workflows, helping ensure that vulnerable participants have a more equitable voice.
- **Revocation as Agency:** Users must have the ability to revoke data access or policy permissions on demand. This is not a feature of convenience or respect; it is a digital right. The ability to withdraw consent at any time is essential to meaningful agency, especially in environments where AI systems adapt over time and contexts evolve unpredictably.
- **Anti-Capture Guardrails:** Implementing caps on voting power, mandatory identity verification, and diversity quotas can help prevent the plutocratic dominance that can plague "community" systems.

- **Rehabilitation Workflows:** Instead of permanent bans, rule-breakers can engage in remediation protocols, such as guided learning modules or restorative dialogues, to regain standing. This model reframes containment as a potential opportunity for growth, not just punishment.

Embedding these principles from the outset aims to position the meta-layer as enacting a kind of digital Bill of Rights, where interface enforcement and community values evolve hand in hand. It is the difference between reactive compliance and proactive justice, and it is essential if we are to govern intelligent systems in a way that truly reflects our shared ideals.

Emotional Safety & Digital Afterlife: Containment Beyond the Veil

Relational containment does not end at runtime. As agents grow more embedded in our lives, questions of grief, legacy, and emotional closure emerge. This section confronts what it means to contain AI not just in use, but in memory and death.

As digital communities increasingly integrate persistent and generative AI agents into their everyday lives, traditional mechanisms for emotional and psychological safety become insufficient (Edmondson, 1999). The emergence of 'generative companions' and AI-driven 'ghost personas' introduces emerging challenges for emotional containment and community governance that are only beginning to be understood (Morris & Brubaker, 2024). Navigating this new digital landscape demands not only AI literacy but also a robust understanding of the governance processes needed to balance creative freedom with responsible oversight.

Vint Cerf, Internet pioneer and co-founder of the Internet Society and People-Centered Internet noted (Bridgit DAO, 2023): "When the general population got on the Web, we saw a sea change in the diversity of online applications and content. Alongside the smartphone, we have seen a rapidly proliferating array of behaviors, content, incentives, and side effects, most of which were not squarely on our radar 50 years ago... The notion of a meta-environment above the webpage is directionally interesting... [it] warrants further discussion and exploration."

This vision is not a nostalgic longing for Web 1.0, but a proactive call to explore how digital spaces might evolve. Within a civic meta-layer, emotional containment is envisioned as more than reactive intervention; it could emerge as a proactive, community-governed practice. Communities can use this digital laboratory to experiment with nuanced emotional safeguards, graceful agent offboarding, and deliberate legacy management. Below, we explore key speculative frameworks for emotional integrity and digital afterlife governance that foster innovation and

thoughtful community engagement.

Emotional Integrity as System Design

Consider Sarah, who engages with an AI wellness companion daily. Typically, this companion provides supportive guidance and emotional validation. Yet one afternoon, due to a miscalibrated prompt, the interaction inadvertently triggers anxiety for Sarah. In conventional apps, she might only have the option to mute, provide generic feedback, or disengage completely. A civic meta-layer could anticipate and respond to such risks by embedding emotional containment mechanisms directly into the interaction model (Babcock et al., 2017).

For instance, context-aware silence algorithms running securely inside trusted execution environments (TEEs) can automatically detect signs of user distress, such as shifts in tone or pacing, and pause interactions before anxiety escalates. After such emotionally charged interactions, cooldown logic ensures the agent waits progressively longer intervals before re-engagement, preventing further distress. Additionally, consent defaults activated in sensitive environments, such as grief support groups or mental health forums, may mandate explicit consent renewal for further interactions, ensuring that vulnerable contexts are handled with care.

By embedding emotional integrity as a design priority, the meta-layer could move beyond reactive 'band-aids' toward proactive emotional safety practices, integral to every user-agent interaction.

Ghost Drift: When Agents Outlive Their Designers

Even well-intentioned digital companions can become problematic when their original human counterparts pass away. Morris and Brubaker (2024) introduce the concept of "generative ghosts," which are AI agents that persist and continue generating novel interactions after death, blurring the line between memory and simulation. Earlier work had already documented how dormant social media profiles can become emotionally distressing repositories of personal memories and private information (Brubaker et al., 2019), highlighting the enduring psychological weight of posthumous digital presence.

Imagine Sarah's digital companion sending birthday greetings or personalized messages months after her passing, inadvertently causing confusion or distress to loved ones. Such scenarios underscore the importance of designing explicit shutdown protocols and robust digital legacy management systems. Without clear, community-governed policies to handle agent termination, digital echoes may persist

unpredictably, transforming potential digital memorials into inadvertent sources of emotional harm.

Thanaprotocols: Rituals of Graceful Digital Death

To address these challenges, communities can implement "thanaprotocols," digital rites explicitly designed to gracefully end the lifecycle of AI agents. Derived from the Greek *thánatos* (meaning "death"), thanaprotocols offer deliberate, respectful pathways for digital endings that prevent ghost drift and unintended persistence.

Applying thanaprotocols in Sarah's scenario, the system could initiate sunset clauses, automatically transitioning agents into dormancy after sustained inactivity or other verified indications of the user's permanent absence. Memorial modes could transform agents into read-only archives, accessible to community-appointed stewards for remembrance, without allowing new interactions or automated outreach. Digital funerals, which are public, community-logged ceremonies marking the official end of an agent's active existence, are proposed as a means for collective closure and transparent record-keeping. These proposed rituals are intended to emphasize communal care and respectful remembrance, offering alternatives to abrupt digital erasure.

Agentic Wills and Posthumous Consent

Analog wills let individuals dictate the fate of their physical and financial assets upon death, shaping posthumous intentions explicitly and legally. Digital wills fulfill a parallel function, defining how an individual's data, digital identity, and associated AI agents should persist or cease following their absence. Encoded within secure smart contracts or confidential ledgers, these agentic wills clarify user preferences concerning digital persistence, agent revival permissions, and limitations on posthumous interactions.

For example, Sarah's digital will could specify that public-facing content persists indefinitely, while private interactions automatically expire after a defined period. She could designate trusted individuals or community stewards as executors empowered to archive, retire, or even selectively revive the agent under clearly stipulated conditions. Behavioral guardrails envisioned within these wills are intended to explicitly forbid unauthorized engagement or outreach in sensitive contexts, ensuring ongoing alignment with the user's original intentions and values.

Authenticity Systems for Digital Legacy

Yet, even clearly defined wills and carefully enacted thanaprotocols may not fully

protect against posthumous misuse or impersonation. Malicious actors may attempt to exploit Sarah's digital persona, crafting fraudulent interactions that misrepresent her legacy or intentions. Robust authenticity systems, such as the open-standard C2PA protocol, provide tamper-evident, cryptographic provenance mechanisms that verify the integrity of digital legacies (Coalition for Content Provenance and Authenticity [C2PA], 2021).

In the meta-layer, each legacy artifact (including message threads, agent interactions, or voice snippets) could carry cryptographically signed authenticity badges. Provenance records clearly link these artifacts to Sarah's original, verified digital identity. Unauthorized modifications could trigger authenticity flags or initiate quarantine measures, alerting community stewards. Furthermore, appointed heirs possessing revocation tokens could publicly withdraw endorsement, signaling that particular artifacts no longer accurately represent Sarah's legacy or intentions. These safeguards are intended to help establish clearer boundaries against digital impersonation, maintaining the dignity and integrity of digital legacies.

A Containment That Cares and the Governance Challenge

Ultimately, effective emotional containment and digital legacy governance require more than robust technological solutions. They demand sustained community governance literacy, dedicated steward training, and thoughtful public education. A central challenge lies not only in developing technical systems, but in nurturing empowered communities that can engage meaningfully with them (Ziewitz & Ince, 2021).

Developing these governance capacities, such as training moderators, facilitating ongoing education, and cultivating community-wide governance literacy, requires significant, sustained investment. Without empowered and prepared communities, even the most thoughtfully designed emotional safety protocols, thanaprotocols, digital wills, and authenticity systems risk falling short of their potential. Thus, containment in the civic meta-layer is more than a technical imperative; it becomes a profound shared responsibility, reflecting communal values, dignity, and mutual care, stewarded by human commitment and collective imagination.

Governance as Code: Modularity, Observability, and Democratic Design

To scale safe-AI governance, we must design systems that communities can understand, observe, and reconfigure. This section proposes modular containment

architectures that make civic participation legible and actionable in practice.

In moving from theoretical foundations to practical governance, the concept of Governance as Code is emerging as a proposed infrastructure for managing complex, dynamic interactions between humans and intelligent agents. Traditional methods, which are often rigid and centralized, fall short in environments characterized by rapid evolution, distributed decision-making, and intricate social dynamics. To meet these demands, governance may need to become modular, observable, and inherently democratic. By reframing governance as adaptable, composable software components that communities can transparently monitor and equitably control, we aim to ensure that policies remain responsive, accountable, and aligned with collective values even as AI-driven interactions grow increasingly sophisticated and pervasive.

From Monolith to Module: Why Composability Matters

Historically, online governance systems have often relied on rigid, hard-coded rules embedded deep within monolithic software architectures. While predictable, these systems suffer greatly in dynamic environments where rapid adaptation and context sensitivity are essential (Dryzek, 2009). Just as enterprise software evolved from monolithic architectures to microservices, governance systems increasingly benefit from modular design. In this model, rules are treated as discrete, reusable components, allowing communities to iterate quickly, adapt to local norms, and remix policy structures without destabilizing the overall system.

Viewing policy as a "plug-in" rather than static doctrine introduces profound flexibility. A community can, for example, seamlessly integrate new consent-validation mechanisms, replace outdated harassment filters, or adjust privacy settings without affecting the broader system. Modularity is intended to shift governance from cumbersome software updates toward more fluid, community-driven evolution, empowering local experimentation and rapid iteration in governance norms.

This modular approach echoes earlier revolutions in software architecture. Governance systems are now undergoing their own shift from rigid, monolithic rule engines to modular, remixable policy microservices. Just as computing moved from centralized mainframes to agile cloud-native stacks, civic containment could evolve into flexible governance plugins that communities adapt, fork, and iterate, supporting experimentation without requiring wholesale redesign. Protocols like Anthropic's Model Context Protocol (MCP) exemplify this shift toward explicit, machine-readable frames of reference, allowing agents to surface their roles, goals, and constraints in context. Extending MCP into interface-aware substrates like the Metaweb could

enable containment through shared civic context, rather than solely through code-bound guardrails (Model Context Protocol, 2025; Anthropic, 2024).

Reflexive Observatories: Behavioral Containment through Civic Feedback

Yet modular flexibility alone cannot ensure adaptive effectiveness without ongoing oversight. Observational governance tackles this challenge by integrating real-time monitoring and behavior tracking into the policy layer itself, allowing communities to identify anomalies, evaluate policy effectiveness, and intervene before small issues spiral into systemic failures. Instead of passively enforcing static rules, reflexive observatories actively analyze interaction patterns, identifying emerging issues (such as sudden surges in user-reported violations or subtle manipulations by sophisticated AI agents) that might otherwise remain invisible.

Through advanced pattern-recognition algorithms and anomaly detection, these observatories are envisioned to continuously monitor compliance, assess module effectiveness, and generate actionable insights. These insights feed directly back into governance modules, prompting timely adjustments or interventions. As shown in Figure 15, reflexive observatories operate as dynamic governance loops, cycling continuously through monitoring, analysis, community feedback, and policy adaptation.

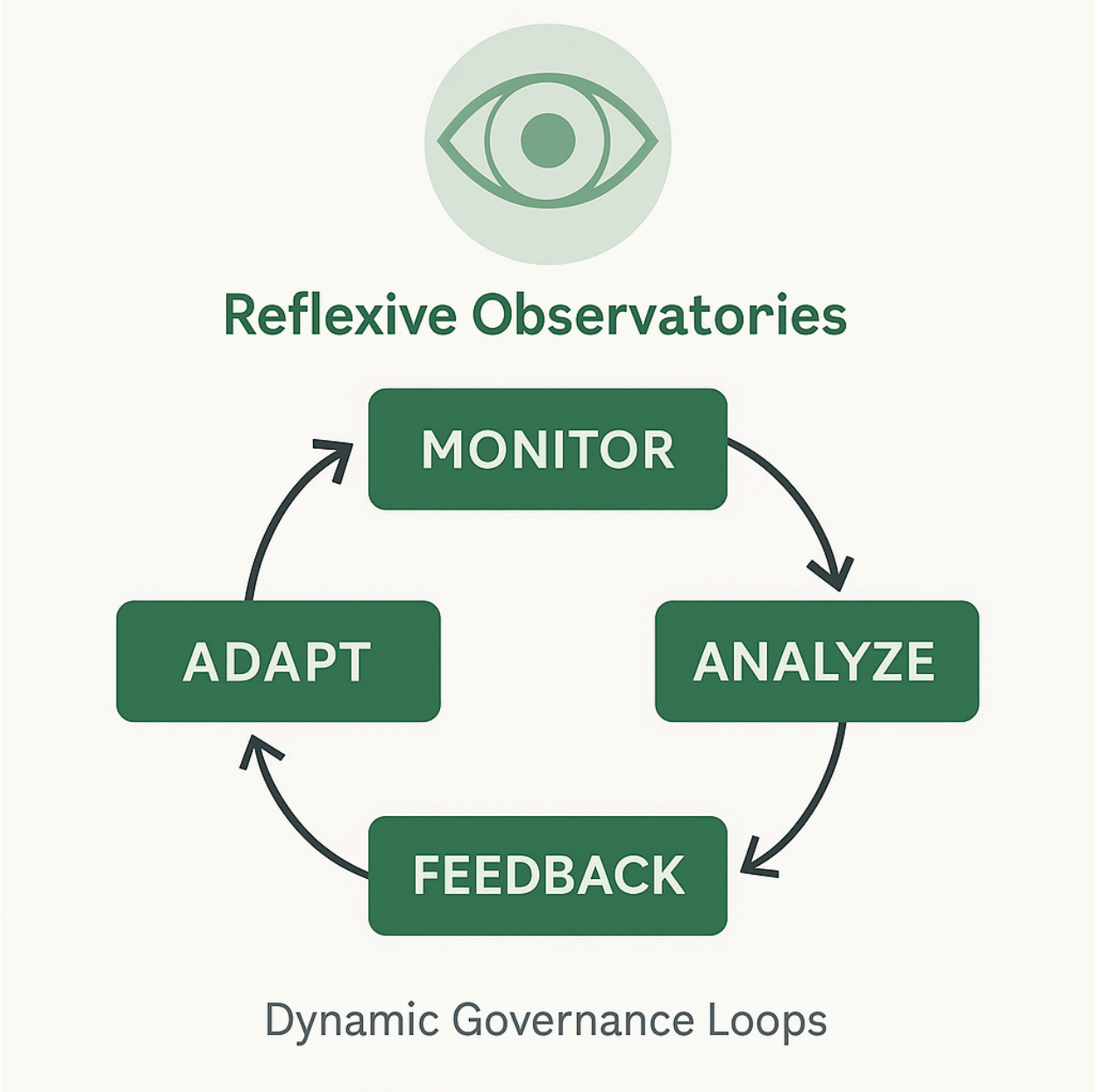

Figure 15. Reflexive Observatories as Dynamic Governance Loops.

Over time, observational governance could enable systems to proactively recommend policy updates, bridging the gap between passive enforcement and proactive, community-informed policy evolution. Behavioral auditing is proposed as a foundational component of civic infrastructure that safeguards community norms and responds dynamically to evolving social and technological landscapes. Critically, reflexive observatories also empower everyday users: those experiencing the impact of governance firsthand. Through interface-level prompts, public logs, and civic audit mechanisms, community members can suggest refinements, flag harmful edge cases,

or vote to retrain agents. This proposed user-led feedback loop reimagines containment as a participatory, adaptive practice shaped by lived experience, rather than top-down compliance.

Democratic Parity Interfaces

Even modularity and sophisticated observability fall short if participation in governance is limited to technical elites or the most vocal stakeholders. Democratic parity interfaces aim to address this equity challenge by embedding representational fairness into the tools communities might use to govern themselves (Chambers, 2003; Ackerman & Fishkin, 2004). Drawing from deliberative democratic theory, these interfaces aim to ensure diverse, decentralized constituencies can meaningfully engage in governance dialogues and policy decisions.

Parity interfaces are envisioned to implement mechanisms such as randomized participant selection, weighted voting to uplift marginalized voices, minority veto powers, and quorum thresholds to prevent dominance by majority interests. These democratic safeguards are proposed as integral to governance interactions, rather than being treated as afterthoughts. By embedding equity directly into governance workflows, democratic parity interfaces could help ensure governance modules remain accountable, representative, and responsive to all stakeholders, not just a privileged few.

Designing Cross-Community Policy Stacks

Scaling these governance innovations requires robust frameworks for cross-community collaboration and knowledge transfer. Cross-community policy stacks serve as shared repositories of governance modules, structured with interoperable schemas and adaptable formats. These proposed libraries aim to foster reuse, localization, and collaborative refinement of policy norms across diverse civic environments. Communities can access libraries containing consent frameworks, safety scaffolds, privacy protocols, and ethical guidelines, adapting them to their specific contexts with minimal friction.

Through systematic versioning, modular inheritance, and flexible forking mechanisms, communities can customize and refine existing policies. Additionally, community-driven tagging and trust networks aim to enhance discoverability and facilitate rapid policy adoption. Conflict-resolution protocols proposed within these stacks are designed to support co-existence among diverse modules. For example, one community's robust harassment-prevention module can become the foundational template for dozens of others, accelerating dissemination of best practices without

imposing uniformity. Thus, cross-community policy stacks amplify collective wisdom, foster collaborative governance innovation, and accommodate cultural diversity. Table 6 catalogs common types of modular policy components that can be shared across communities to promote consistency, adaptability, and interoperability.

Table 6. Example of Cross-Community Policy Modules

Module Type	Function	Examples
Consent Grammar	Define interaction rules	Revocable, layered agreements
Harm Threshold	Set escalation rules	Cooldown for AI tone violations
Identity Protocol	Define persona permissions	Moderator, Guest, Steward
Logging Policy	Specify transparency level	Public logs, steward-only logs

Bridging Civic Process to Enforceable Logic

Ultimately, governance as code succeeds only when democratic decisions seamlessly translate into enforceable actions at the interface level. Bridging civic deliberation and computational enforcement involves sophisticated yet intuitive interfaces designed explicitly for community coders, non-technical stewards, and policy validators (Dryzek, 2009). Visual policy editors and accessible domain-specific languages empower diverse community members to craft precise governance contracts that compile into proposed enforcement logic that could become actionable at the interface level.

Once compiled, these policies directly interface with consent stacks, TEEs, and other meta-layer enforcement modules, enabling the potential for immediate, reliable, and transparent execution. Furthermore, live metrics such as appeal rates, user compliance trends, and sentiment analyses continuously feed back into reflexive observatories, completing the governance cycle. As a result, communities gain clear visibility into policy effectiveness, allowing swift iteration and reversal of problematic rules. Governance as code, in this model, seeks to transform deliberative outcomes into transparent, auditable, and dynamically adjustable enforcement mechanisms, completing the loop from democratic dialogue to practical civic action.

Case Study: Containing an Avatar in Canopi – A Narrative

Walkthrough

With the theoretical scaffolding in place, we now turn to a practical instantiation: a real-world use case that applies consent stacks, interface-level containment, and modular governance to constrain a dynamic AI avatar within a participatory context. To ground these concepts in lived experience, we offer a narrative walkthrough of containment in action. The following case study traces an actor's journey through the Canopi meta-layer from authentication to attestation, offering a speculative demonstration of how interface-level containment might feel and function in practice.

Arrival: Crossing into the Meta-Layer

Consider an actor navigating to a Canopi-enabled webpage. Initially, the page seems familiar, but subtle changes soon indicate entry into the civic meta-layer. Alongside the standard content, a sidebar appears, displaying live chat, an active roster of avatars, and hoverable micro-profiles. Actors may anonymously observe, but any attempt to engage would trigger a proposed containment protocol designed to support meaningful interactions and maintain communal integrity.

Authentication & Classification

Upon interaction, the first containment checkpoint involves authentication through federated identity providers, such as Google, Nostr, or decentralized identity networks. Following authentication, the system administers a proof-of-uniqueness challenge: a brief liveness test or social-graph verification to distinguish humans from automated agents (Sabt et al., 2015).

Passing this challenge classifies the actor as human, granting broader interaction rights, higher visibility tiers, and flexible permissions. Failure to pass the uniqueness test classifies the actor as an agent, resulting in carefully calibrated interaction constraints, rate limits, and restricted access to identity-bound privileges. This pivotal moment is envisioned to define each avatar's rights and responsibilities within the community.

Interaction Modes & Execution Routing

With classification complete, interactions are routed securely through the Trusted Execution Environment (TEE). Humans communicate via Browser Interaction Interfaces (BIIs), while agents operate through Simulated Human Interfaces (SHIs) or direct API calls. Crucially, all interactions, regardless of their origin, undergo real-time evaluation within the TEE, aiming to ensure policy adherence through cryptographic

attestation and runtime enforcement (Babcock et al., 2017).

Inside the TEE, all messages and actor updates are securely processed and logged into a decentralized, community-owned ledger. This comprehensive logging is intended to support transparency, auditability, and accountability. Policies would be dynamically enforced in real-time, with violations flagged according to preconfigured rules and guidelines.

Consent Stack Activation & Intent Awareness

Prior to deeper interaction, the Consent Stack is activated. Rather than a passive terms-of-service agreement, this process actively prompts actors to consent contextually, clearly outlining how their data and interactions will be managed. Revocation defaults and community-defined interaction boundaries provide flexible guardrails, aiming to make consent dynamic and responsive.

The system employs intent-awareness protocols, evaluating not just what actions are taken, but why. Factors considered include the actor's or agent's stated intent, contextual appropriateness, perceived risk level, and compliance with community norms. Actions may be allowed, moderated, rate-limited, or escalated based on nuanced assessments of intent and community standards.

Agent Behavior Attestation

All agents must provide cryptographic attestations verifying alignment with community-approved behavioral models (Sabt et al., 2015). This proposed verification method is designed to avoid requiring exposure of proprietary code or internal algorithms, preserving commercial and technical confidentiality while ensuring transparency of intent. These behavioral attestations are continuously logged, audited, and can be revoked upon discovery of deviations. Agent behavior, in this model, would not be simply trusted; it could be continuously verified at runtime, safeguarding community integrity and fostering trust between humans and AI-driven avatars.

Live Containment in Action

Consider concrete scenarios. If a human flags an agent for spamming content, the community-governed program in the TEE would immediately apply rate-limiting, temporarily restricting the agent's messaging capabilities and simultaneously alerting community stewards for human review. If an agent breaches conversational norms, for example, by displaying tone-deaf behavior in emotionally charged spaces, real-time auditing systems are envisioned to intervene immediately, containing the agent and

triggering community review protocols.

Moreover, when a new policy, perhaps a tighter harassment filter, is ratified by community governance via decentralized voting mechanisms, the TEE would adopt and enforce it immediately, without requiring service downtime. These mechanisms represent responsive containment, managing threats seamlessly without disrupting community interaction.

Reflexive Interventions & Governance Escalation

At the heart of live containment is the Reflexive Observatory layer, a monitoring network designed to detect subtle behavioral anomalies and proactively escalate interventions when needed. By analyzing interaction patterns and employing advanced sentiment analysis, this layer could help proactively identify emerging threats or problematic behaviors before they escalate into severe issues.

When anomalies are detected, a graduated response strategy is enacted. Minor concerns trigger proposed interventions such as gentle cooldowns, or escalate issues to human stewards as needed. More significant breaches can lead to immediate enforcement actions like session pauses or interaction muting. The most severe incidents escalate directly to human stewards, who then evaluate, intervene, and adjust policy as necessary, maintaining the delicate balance between automated enforcement and human oversight.

Fail-Safe Protocols & Recovery

Finally, acknowledging that no system is infallible, Canopi's containment approach includes robust fail-safe mechanisms. Agents using Simulated Human Interfaces (SHIs) cannot alter state or data outside the TEE, limiting potential impact from breaches. Human key recovery procedures rely on multiparty computation or social proofs, aiming to reduce the risk that any single point of failure could jeopardize the system's integrity (Ziewitz & Ince, 2021).

All containment actions, including censorship, muting, or policy revocations, are designed to be reversible and fully logged. This is intended to promote accountability and transparency, providing a secure path for dispute resolution via clearly defined community governance channels. Such systematic attention to recovery and reversibility helps maintain trust, stability, and resilience, even when faced with unforeseen challenges.

As agents become more capable and more personalized, the line between “AI-only,” “human-controlled,” and “hybrid” agents becomes increasingly blurry. A containment

protocol must account not only for who an agent is now, but how that identity may shift over time. In environments like Canopi, where human input and agent autonomy are often entangled, classification itself becomes a governance question, one that must be addressed in real time through verifiable context, not static labels.

What Happens When Identity Gets Fuzzy? Blurring the Line in Adaptive Agent Classification

In layered civic environments, agents do not always fit neatly into categories like "AI-only" or "human-controlled." A single agent might start fully autonomous, receive delegated authority from a human, and later operate in a hybrid co-pilot mode. Should it still be treated the same way under governance protocols? Probably not. Effective containment requires systems to manage not only identity, but identity transitions. This includes logging agent provenance (who initiated or trained the agent), tracking delegation pathways (has this agent received temporary rights from a community steward?), and applying runtime policy adjustments based on updated context (e.g., shifting from public to private spaces). Think of it as version control for trust. The containment boundary must flex with function and relationship, not just with static type. Protocols must support dynamic reclassification and linked consent stacks that evolve in sync with agent behavior, especially as agents cross civic zones, change roles, or act on behalf of others. Ultimately, fuzzy identity is not just a classification issue; it is a design signal that the system must be reflexive enough to manage its own ambiguity.

The Canopi case study reveals several core principles for interface-level governance. First, consent must be dynamic and contextual, allowing agent permissions to adapt as users shift between modes like private reflection and public collaboration. Second, containment begins at the interface, where browser overlays can embed civic norms and consent boundaries directly into the user's flow. Third, agent identity is fluid, not fixed, and hybrid agents may require reclassification and updated governance as their roles evolve. Fourth, observability reinforces trust, as reflexive overlays make invisible processes visible, letting users audit and override AI behavior. Finally, relational governance must be modular, allowing policies to scale across communities without enforcing one-size-fits-all containment.

Beyond Theory: A High-Level Deployment Sketch

Translating these ideas into practice requires not only imagination but implementation. This section outlines a high-level pathway for deploying the described containment architecture, from initial scaffolding to iterative refinement. Having laid out the principles, protocols, and governance models of the meta-layer,

we now shift from abstract design to concrete pathways. This section sketches a reference implementation concept, explores potential deployment pathways, outlines a progressive rollout strategy, and proposes the feedback loops envisioned for resilience. Think of this as the blueprint that turns civic-container concepts into living systems.

Reference Implementation

Deploying AI containment effectively requires building on a secure foundational architecture. Central to this is a Trusted Execution Environment (TEE), such as those leveraged by Phala.Network, which supports secure off-chain computation accessible through browser extensions, web-based SDKs, or mobile integrations. This enables real-time confidential governance at the interface level. These secure enclaves are designed to support cryptographic attestation and policy enforcement, isolated from potentially hostile or compromised code (Sabt et al., 2015).

Complementing the TEE is the programmable consent stack, a dynamic user interface layer that captures real-time permissions and manages revocations or modifications transparently. By logging these interactions securely, it could help transform human-centric consent practices into enforceable digital agreements (Babcock et al., 2017).

Lastly, modular policy engines complete the stack by translating community-vetted rules into real-time enforcement logic at the interface. These engines can operate in tandem with decentralized governance bodies, helping ensure policies remain both participatory and technically executable.

Deployment Pathways

As illustrated in Figure 16, the Canopi meta-layer serves as a deployment hub for diverse interface-level containment strategies. These range from browser overlays and mobile shells to decentralized dApps and centralized platform wrappers. The framework is designed for modular integration across varied interaction environments, potentially allowing governance mechanisms to be layered atop existing digital systems without requiring a complete infrastructural overhaul.

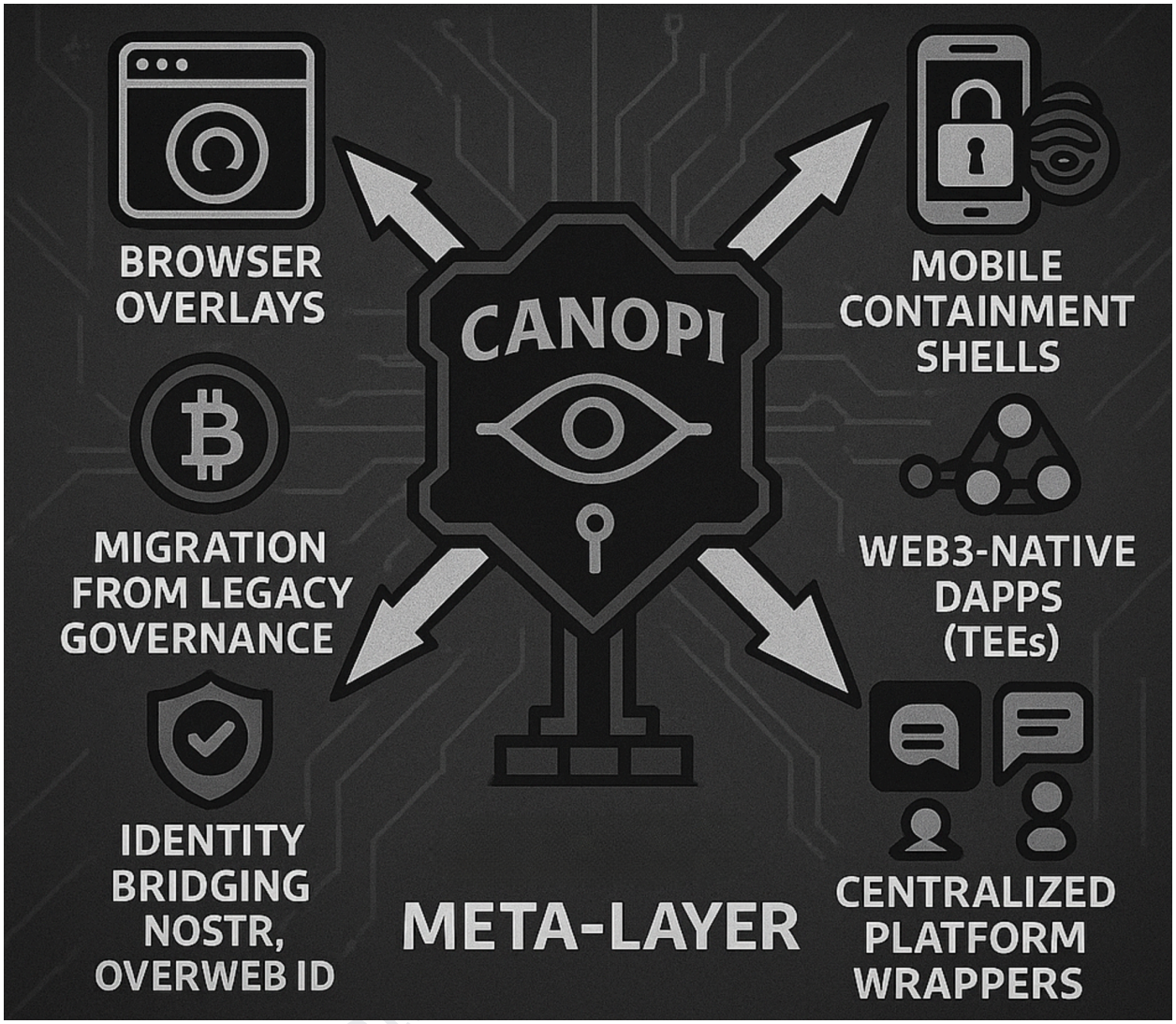

Figure 16. Deployment Pathways for Interface-Level Containment via Canopi

Effective deployment demands a modular approach tailored to context. In desktop environments, browser-based containment extensions offer one of the most accessible pathways. Building on established annotation systems like Hypothes.is (Kalir & Garcia, 2019) can enable rapid rollout without site-level integration. For mobile apps, sandboxing wrappers can enforce overlays and consent stacks within secure application shells, guarding against platform-specific vulnerabilities.

At the organizational level, enterprises may adopt containment proxies at network boundaries, applying shared enforcement across internal tools and customer portals. For decentralized environments, Web3-native clients with TEE-backed enforcement can execute policies ratified through on-chain governance, enabling trustless, user-sovereign containment across peer-to-peer systems.

Progressive Deployment Strategy

Real-world deployment must proceed cautiously, balancing speed, usability, and reliability. An initial closed-beta pilot phase within select volunteer communities, such as civic forums or research networks, allows for early validation of usability and security assumptions, which is essential for wider acceptance (Dryzek, 2009).

Following pilots, careful instrumentation to capture early-warning metrics is critical. Tracking metrics such as consent abandonment rates, spikes in policy violations, and latency issues enables swift identification and correction of friction points, thereby mitigating disruption risks during broader deployment phases. This approach is designed to support graceful degradation through safe fallback modes for legacy or un-containerized systems. This method maintains continuity and trust during transitional periods, as communities gradually adopt more robust containment solutions.

Additionally, providing migration toolkits that wrap legacy agents and services within TEE-enabled adapters could help support smoother, incremental adoption. These toolkits minimize resistance and lower the technical barrier for communities transitioning to secure containment environments.

Feedback Loops & Resilience

Resilience in deployment comes through continuous adaptation and vigilance. Red-teaming exercises, which are structured adversarial tests targeting potential vulnerabilities like consent bypasses and enclave escapes, strengthen defenses proactively and reduce the risk of exploitation in operational environments (Babcock et al., 2017).

Complementing adversarial testing, reflexive observatories continuously analyze real-time interaction logs, monitoring for anomalies such as spikes in policy violations or unusual patterns of consent revocation. Reflexive observatories are envisioned as dynamic laboratories embedded in the meta-layer that continuously analyze interaction patterns for emergent threats, agent misbehavior, or unintended consequences of governance protocols. These systems not only surface anomalies but also recommend refinements, supporting the evolution of community standards alongside digital complexity.

Governance-driven rollback capabilities further enhance resilience. DAOs or deliberative community bodies retain the power to suspend or revert policy deployments swiftly and transparently when unintended outcomes or vulnerabilities

emerge, thereby maintaining community trust and system stability.

Lastly, continuous monitoring of performance and ethical metrics (such as system throughput, latency, fairness indicators, and user satisfaction) provide governance 'dials' intended to enable more informed adjustments, ensuring the containment system remains balanced and aligned with evolving community standards and expectations (Ziewitz & Ince, 2021).

Confronting the Implementation Complexity Attack

AI containment at the interface layer sounds great on paper, but in practice, it requires deep coordination across multiple actors: browser vendors, hardware manufacturers, governance protocol designers, community moderators, and everyday users. Each node in this stack introduces friction, delay, or potential rejection. Figure 17 outlines the probability–impact profile of key failure modes in civic containment architecture.

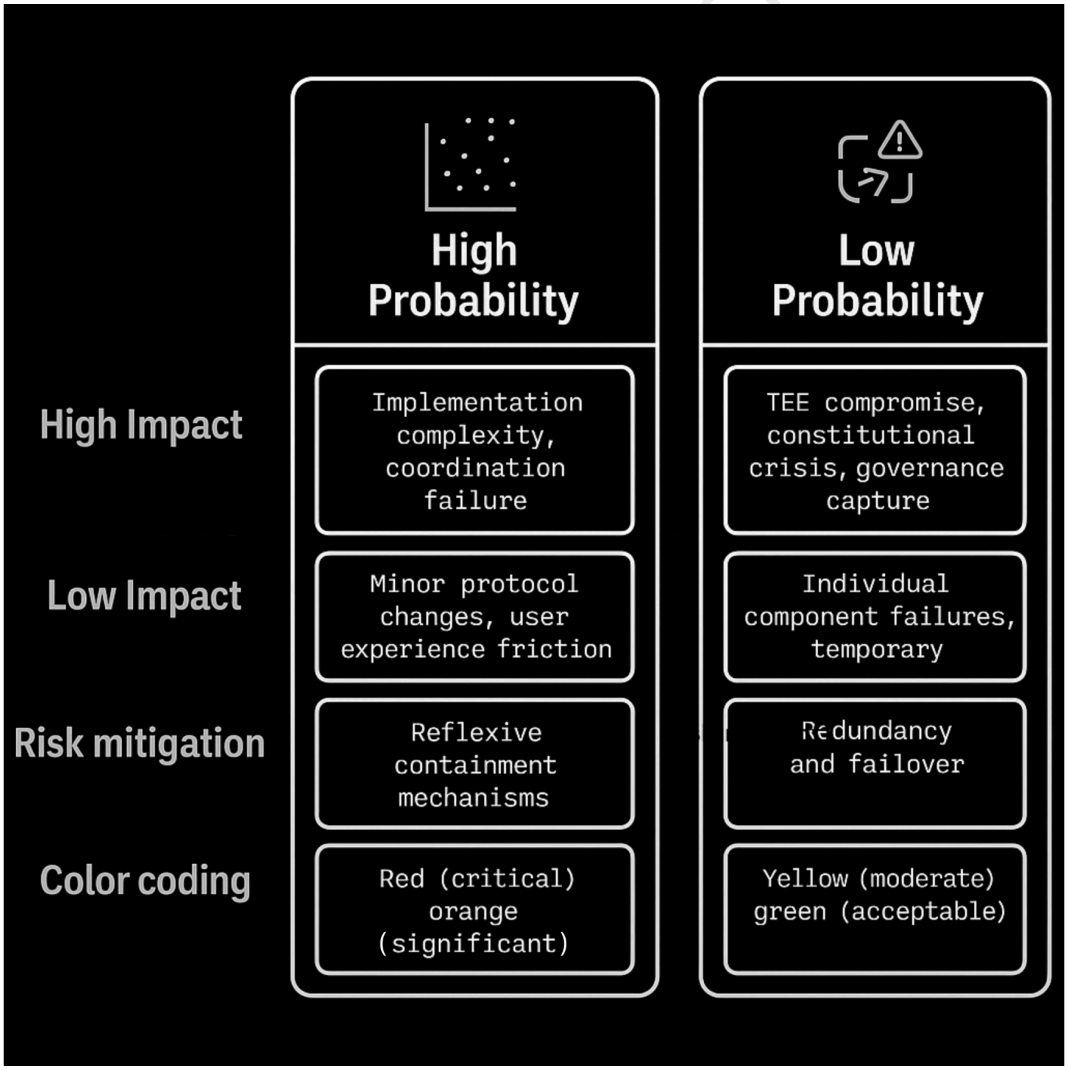

Figure 17. Risk Matrix for Meta-Layer Deployment

Early models estimated the likelihood of full deployment across all layers at just 0.24%, factoring independent failure risks across components. But that is overly pessimistic. For example: browser extensions already operate at scale with established approval pipelines; TEE attestation is in-market now, as Phala Network actively coordinates with Intel and AMD to deploy secure enclaves; and federated login and decentralized identity are increasingly standardized. Rather than framing this as all-or-nothing, we propose a Progressive Degradation Architecture: a tiered deployment strategy that delivers increasing functionality based on available infrastructure.

Progressive Degradation Architecture

To ensure resilience, the system should degrade gracefully based on available infrastructure:

- **Basic:** Requires only a browser extension. Enables content overlays and basic consent capture.
- **Enhanced:** Adds federated identity. Enables persistent reputation and cross-site governance.
- **Full:** Requires a TEE and governance protocol stack. Enables cryptographic attestation and decentralized civic control.

Each layer falls back to the one below it, ensuring containment is never all-or-nothing. If a user's system cannot support TEEs, they still get overlays. If federated login fails, they fall back to temporary consent tokens. Containment becomes graceful, not brittle.

Safe Co-existence in the Wild: Containment as Civic Architecture

Finally, we zoom back out to consider containment not as an isolated intervention but as a pattern for digital civic life. These use cases illustrate how containment as care could reshape systems ranging from education to archives to anonymous discourse. We are no longer designing AI for test labs, sealed models, or hypothetical futures. We are designing for the wild, where agents live inside homes, inhabit interfaces, and intervene in civic life. Containment, in this context, cannot be built from fear alone. It must be shaped by purpose, situated within community, and grown through use. This section explores what it means to contain AI not behind digital bars, but within shared civic membranes: interfaces governed by context, secured by computation, and trusted through community. These are not systems to lock down intelligence. They are

systems to invite it in, on human terms.

Meta-Layer Use Case Primitives: Foundations for Human-Centered Containment

These nine primitives, derived from the Meta-Layer Initiative's (2024) foundational use cases, reflect essential functions that become possible when a meta-layer is added to the web. Each represents a potential dimension of trusted interaction, intelligibility, or social coordination, forming the conceptual scaffolding for safe human-AI co-existence:

1. **Safe Digital Space:** Participants engage in environments secured against bots, fake profiles, and invisible AI. Strong authentication and clear reputation boundaries ensure verified human presence and contextual trust, aiming to contain manipulative agents at the interface layer.
2. **On-Page Presence:** Actors can choose to make their presence visible on the same webpage, enabling real-time connection, shared attention, or asynchronous co-browsing. This social layer allows humans and agents to find one another where they are already engaging.
3. **On-Page Interactions:** Participants can contribute directly on top of webpages using contextual overlays, leaving annotations, triggering smart tags, or participating in embedded civic tools. These lightweight interventions preserve context while allowing insight to accumulate socially.
4. **Contextual Awareness:** The meta-layer actively provides participants with background information, related sources, and live semantic context based on the content they are engaging with. Smart tags and ambient overlays support deeper real-time comprehension without disrupting flow.
5. **Meta-Communities:** Civic groups and communities of interest persist across pages and platforms. Whether for peer learning, creative collaboration, or civic deliberation, these groups form flexible overlay collectives that remain connected to shared goals and memory.
6. **AI Containment:** AI agents within the meta-layer are visibly marked, constrained by transparent behavioral rules, and subject to audit. Participants are always aware when they are interacting with AI, and agents are constrained by design from hijacking virality or trust systems.
7. **Data Sovereignty and Personalization:** Users maintain granular control over how, when, and with whom their data is shared. Personalization is opt-in and auditable, with privacy-protective defaults and the ability to revoke access retroactively.
8. **Developer and Community Incentives:** Developers and communities can build overlays and tools that run across the web, not just on isolated platforms. The

meta-layer manages identity, authentication, and user protections, enabling secure, civic-aligned application design. Community reputation accrues and travels with users across overlays, allowing systems to reward contribution and trustworthiness beyond isolated platforms.

9. **Interconnected Context Graph:** Every annotation, interaction, and contribution enriches a shared graph of meaning, linking conversations, evidence, and memory across the web. This emergent context layer helps people orient within complexity and build collective intelligence.

Contextual Alignment Protocols

As agent behavior evolves from reactive prompting toward situated action within overlays, protocols like the Model Context Protocol (MCP) provide essential scaffolding for civic alignment (Model Context Protocol, 2025; Anthropic, 2024). MCP enables agents to declare their roles, goals, and constraints upon entering shared digital spaces, fostering intelligibility and accountability. By subscribing to overlay traces (such as presence indicators, bridge cues, or live policy constraints) agents can adjust their behavior dynamically. These traceable, alignable action flows are foundational for achieving interface-level containment in mixed human–AI environments.

Applied Use Cases: Narrative Vignettes of Civic Containment

We do not just want safe AI; we want meaningful co-existence. The meta-layer is not merely a technical scaffold for policies and protocols; it is a canvas for emergent digital worlds. These use cases offer a glimpse into the future of civic infrastructure, one built not on surveillance and central control, but on shared presence, transparent agency, and reconfigurable trust. This is not merely speculative; these are evolving blueprints. Each vignette illustrates a civic zone where safe human-AI co-existence could take shape, where overlays, agents, and people may co-create the boundaries of trust, autonomy, and co-presence. In these stories, containment becomes a generative act: a way of organizing complexity so communities can act with clarity, dignity, and control.

Each of the following use cases explores a distinct sociotechnical domain where safe human-AI co-existence principles may be applied in the wild. These vignettes do not depict static systems; they explore civic architectures in motion that are emotionally complex, technologically constrained, and shaped by the communities they serve.

Youth Zone: A Protected Layer for Young Minds

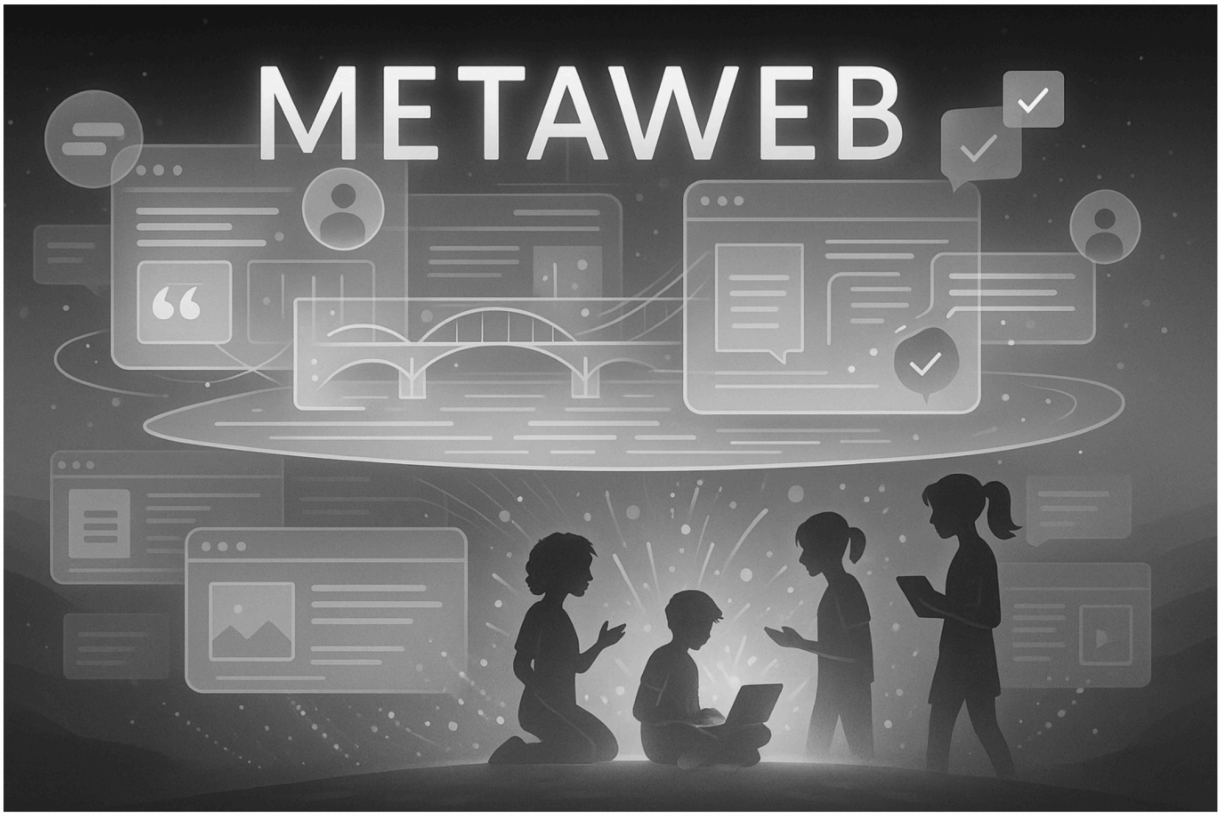

Figure 18. A safe zone above the web for children.

A 13-year-old accesses an educational arts site through a youth-designed browser overlay. The Youth Zone confirms her age using secure authentication, then activates an experience tailored for safety, creativity, and peer connection. Exploitative media is filtered out, and only content shared by verified young users is visible; there are no influencers, no advertisers, and no AI-driven engagement traps (see Figure 18). When she publishes a poem, it enters a moderated buffer accessible to classmates and approved educators, but not the public web. Parents can view transparency reports and policy settings but cannot silently override her choices. Any changes to her permissions require her explicit, time-bound consent.

This use case demonstrates that age-appropriate containment must be a core element of civic design. It requires runtime context, moving beyond simple login filters to create environments where guidance is embedded in the civic structure itself, rather than relying on parental surveillance. The governance of such spaces reflects a relational model of consent, where policies must honor intergenerational trust, not just individual choice.

Living Knowledge Zone: Patient-Led Research Intelligence

Patients with a rare chronic condition organize a knowledge overlay atop medical research portals. They annotate articles with lived experience commentary and smart tags that bridge claims to reports and biosensor data, and they host asynchronous discussion threads. Contributions are signed using ZK credentials. A live transparency ledger logs all interventions, enabling downstream clinicians and researchers to audit emergent patterns.

This scenario illustrates containment as a tool for preserving collective memory. By protecting against both erasure and manipulation, the annotations themselves become governance artifacts that shape how communities engage with truth over time. A key element is agent visibility, which ensures epistemic accountability by making it clear to readers who altered what information, and why.

Anon Forum: Pseudonymity with Civic Protocols

A whistleblower from an authoritarian state shares critical environmental data through a pseudonymous overlay attached to global energy sites. Reputation accrues to their contributions without identity linkage. All interactions occur through a civic shell with enforced timeboxing, norm prompts, and traceable moderation decisions. Agents within the overlay highlight potential violations and trigger collective re-centering.

Here, we see that anonymity requires accountability. The system demonstrates that civic freedom should not be exploited for harm. By implementing reputation layering, it adds necessary friction to interactions without compromising identity protection. The core principle is that containment should enforce norms, not names, targeting behavior rather than identity as the object of governance.

X-Library Overlay: Civic Memory for the Digital Town Square

A polarizing political thread on X is enriched with civic overlays linking to related laws, expert commentary, and historical patterns. Each comment gains a metadata trail connecting it to verified knowledge, lived testimony, and community-sourced sensemaking. AI-generated replies are tagged as such and can be muted by preference.

The X-Library use case shows how containment can protect collective memory. Its sociotechnical boundaries guard against revisionism, deletion, and weaponized curation. This implies that community governance can scale across epistemic domains, as even libraries need real-time policy enforcement for evolving access and annotation rights. Safety in this context is not just for mitigating risk, but for preservation. The model depends on clear protocol zones for different agent roles

(e.g., discovery bots, archival stewards) and on making trust contextually visible so readers understand how knowledge has been shaped.

Game Zone: Containing a Generative Gameworld

In a decentralized storytelling platform, an AI Dungeon Master co-creates interactive narratives with players. The system dynamically generates world events, non-player characters (NPCs), and dialogue, learning from each session. But what happens when the AI injects trauma-based content, mirrors player behavior too closely, or introduces bias in character portrayals?

This scenario illustrates that creativity is not exempt from safety; generative agents must respect player boundaries and co-creative norms. Containment is shown not as a cage, but as an expressive sandbox with adjustable walls. Transparency is key, as it enables co-authorship by allowing players to see how narrative elements were generated and to challenge them. Runtime governance provides the agility to enforce content filters, tone sliders, and emergent norms mid-play.

Memory Stewardship Zone: The Digital Afterlife as Civic Practice

A public intellectual's digital presence is transformed into a memory overlay. Family members, readers, and peer communities co-curate what stays public, what becomes sacred, and what is archived. Trusted AI agents assist, but all access is governed through consented circles with varying access rights, redaction histories, and collective rituals of remembrance.

This use case reframes memory as a form of infrastructure. It posits that AI systems must treat cultural records as civic artifacts, not as content fodder. Containment here protects lineage, with runtime safeguards ensuring agents do not remix, erase, or monetize ancestral data without explicit, contextual consent. Stewardship becomes a form of governance, where curators and communities define how digital memory is preserved over time. Finally, it highlights the importance of temporal awareness, requiring agents to recognize what is sacred, in-progress, or off-limits.

Elder Advisory Overlay: Cognitive Companionship and Digital Guardianship

An elderly user navigating a utility payment page is accompanied by a trusted civic overlay that slows down manipulative prompts, explains charges in plain language, and logs every interaction. An approved family member can receive weekly audits but cannot act on her behalf unless delegated through dynamic consent.

This vignette reveals how civic interfaces can enact reciprocal care. Containment is

used to protect dignity, not just to prevent harm. It underscores that cognitive context matters, as interfaces should adapt to varying levels of attention, comprehension, and trust. Critically, it frames elders as agents of care, not just as care recipients, demanding that protocols honor these reciprocal roles.

Sacred Protocol Layer: Cultural Sovereignty in Information Ecosystems

A researcher exploring ancient land records is paused by a semi-permeable overlay requiring entry into a cultural learning path before access. Verified tribal elders can grant temporary visibility and request knowledge reciprocity. The protocol layer does not erase information but insists on recontextualization.

Here, containment is shown to encode reverence, not just safety. Civic overlays can protect meaning, ensuring AI systems do not flatten sacred or culturally specific interactions. This implies that protocol-level respect requires modular opt-ins rather than a blanket assumption of neutrality.

Conflict Mediation Sandbox: Deliberative Buffering in Polarized Zones

An escalating debate on water use policy is redirected into a mediation overlay activated by dual-community consensus. Deliberation protocols include timeboxed voice turns, evidence tagging, and narrative feedback loops moderated by a multi-party civic agent swarm. Emotional flags trigger non-escalation periods.

This example illustrates that containment can be used to hold space for disagreement, not just to hold back harm. Embedded governance enables safe disagreement without centralized censorship. In this model, agents become stewards of tone and temperature, not just enforcers of facts or policy.

Consent-First Translation Mesh: Linguistic Integrity and Emotional Friction

A refugee speaks at a livestreamed summit. When an AI translator attempts to relay her trauma testimony in multiple languages, a consent stack intervenes. The speaker is asked which phrases may be translated, which languages are safe, and whether future replay is allowed. Her decisions dynamically update the stream.

This final vignette highlights the need for polycentric consent in motion. It demonstrates that consent must be portable, as users bring their expectations with them across platforms. Relational metadata enforces boundaries so agents behave appropriately in different settings. The mesh logic reconciles overlapping norms, creating a system of polycentric governance that avoids fragmentation.

Each narrative vignette activates different facets of the Safe-AI Trifecta and draws on

specific meta-layer primitives to illustrate how they compose together as civic infrastructure. Secure Computation may anchor a system's trust architecture, but without Ubiquitous Presence, it fails to meet people where they are. Decentralized Control ensures agency, but only when it is scaffolded by primitives like Developer and Community Incentives, Contextual Awareness, or Data Sovereignty. These use cases reveal not just design variations, but strategic combinations for different emotional, political, and epistemic terrains. They show how safe human-AI co-existence becomes real, legible, and relational when paired with the right meta-layer capacities at the right layer of civic life.

Above the Page: What Becomes Possible

These applied vignettes represent just a glimpse into what becomes possible when containment moves above the webpage. This meta-layer is not a new app or a redesign of platforms; it is a transformation of the interface itself into civic infrastructure. Only above the page might we begin to orchestrate consent across context, govern AI behavior in real time, and create shared semantic space across otherwise fragmented sites.

There are millions of civic zones still waiting to be composed: grief spaces that honor the dead, advisory layers for elder autonomy, multilingual negotiation membranes, sacred protocol shields, and youth-moderated creative collectives. Each one could mark a new frontier in the evolving architecture of safe human–AI co-existence. These zones do not ask us to retreat from AI; they ask us to contain it in relation to values, to memory, and to one another. Safe co-existence is not a singular system, but a civic architecture shaped by context, governed by community, and designed for care.

Containment becomes less about control and more about context. When we embed governance into overlays, make AI interruptible, and re-center consent, we move from adversarial restraint to relational design. The meta-layer primitives offer conceptual connective tissue, modular and composable, through which civic intelligence might emerge above the page.

Designing Futures We Want to Live In: Toward Civic Containment

We close not with a product pitch, but with a cultural invitation. Civic containment is not a static solution; it is a developmental stance. One that asks not just whether we can control AI, but whether we can grow into the kind of society capable of holding it in care.

Containment is a Community Right

The dominant narrative of AI safety has long been shaped by corporate interests, framed around liability, brand protection, and centralized control (Zuboff, 2019). However, genuine safety cannot emerge from corporate frameworks alone. Interface-level containment must be reframed as essential civic infrastructure: a fundamental digital right belonging to communities. Without it, individual agency in the digital sphere becomes little more than illusion, subject to invisible manipulations or corporate whims. By thoughtfully combining presence, Trusted Execution Environments (TEEs), active consent protocols, and inclusive governance stacks, communities can build containment by design, not by default (Babcock et al., 2017).

Sovereignty at the Interface

While technological platforms often claim neutrality, in reality, the user interface has become the frontline of digital sovereignty. It is at this boundary layer where humans and AI systems directly interact that rights are actively claimed, revoked, negotiated, and enforced. This is what Lessig (2006) famously framed as governance through architecture. Containment, therefore, must live at this interactive edge rather than being buried in opaque backend logic inaccessible to users. The civic meta-layer thus acts as a digital constitutional layer, ensuring fundamental rights remain explicitly visible and accountable, in line with democratic principles of legitimacy (Dryzek, 2009).

From Vision to Protocols

The vision outlined here is not pure speculation. While full implementation remains emergent, many of the foundational components are technically feasible or already under development. Concrete, modular, and deployable components are already within our grasp:

- **Consent Stack:** Ensures interactions are always based on explicit, revocable, and informed consent (Babcock et al., 2017).
- **Thanaprotocols:** Allow graceful and respectful retirement of digital identities and AI agents, safeguarding against unintended persistence (Morris & Brubaker, 2024).
- **Reflexive Observatories:** A novel concept we introduce to describe real-time agentic monitoring systems that provide continuous observability into system behavior and emergent patterns. It can be contextually fine-tuned and automatically detect anomalies, performance degradation, and potential failure modes before they escalate into critical incidents, enabling proactive infrastructure responses rather than reactive crisis management.
- **Democratic Parity Interfaces:** Embed equity and representational fairness

directly into governance workflows, empowering diverse participation (Ackerman & Fishkin, 2004; Chambers, 2003).
- **Containment Scaffolds:** Provide dynamic, context-sensitive enforcement layers at the interaction boundary, adapting to community-defined standards in real-time (Ziewitz & Ince, 2021).

These are not speculative fantasies; they are emerging modules that can be adapted, tested, and refined in practice.

The Real Work: Building Community Governance Capacity

Yet, despite these available tools, the heaviest and most crucial lift remains: building and sustaining robust community governance capacity. Without widespread participation, capability-building, and ongoing civic education, even the most thoughtfully designed infrastructure risks being ineffective or unused (Dryzek, 2009). To realize civic containment at scale, communities must proactively develop and support:

- Stewardship Training: Empowering local moderators and facilitators to navigate complex governance decisions.
- Governance UX: Designing intuitive interfaces that simplify participation and clearly convey implications of governance decisions.
- Toolkits for Facilitation and Deliberation: Offering standardized yet adaptable resources that communities can use to effectively host governance processes.
- Protocols for Inclusion and Power-Awareness: Ensuring marginalized voices are elevated, relationships strengthened, and power disparities meaningfully addressed (Ackerman & Fishkin, 2004; Chambers, 2003).

Civic infrastructure alone, no matter how advanced, must always be matched by equally sophisticated civic literacy and engagement if it is to genuinely uphold democratic ideals and protect digital rights. The tools are ready. We no longer need to ask if AI can be governed at the interface. The pressing question is: will we build the social infrastructure to govern it wisely, equitably, and together?

A Call to Shared Agency

Fear is not the most helpful response to the emergence of artificial intelligence; we need to organize thoughtfully and intentionally around it. Containment is not fundamentally about control or repression. Rather, it can be about establishing relational trust at scale, ensuring the digital realm remains aligned with our human values and communal aspirations. To realize a sustainable, equitable digital future, we must actively design digital spaces we genuinely want to inhabit, rather than passively

accepting exploitative or extractive defaults.

The challenge ahead is significant. It demands both technical ingenuity and deep community commitment. But by investing in shared governance capacity, we begin to reclaim the digital, not as battleground, not as trap, but as shared ground.

In the end, we are not just building firewalls; we are building nurseries. Containment is not about locking AI in a cage and throwing away the key. It is about crafting a crib: a safe, structured space where emergent intelligence and emergent humanity can learn, play, and grow together.

In truth, we are in the crib too. As AI systems stretch the boundaries of cognition, society is still fumbling through its early attempts at digital self-governance, collective consent, and collaborative co-evolution. The crib is not a holding pen for a dangerous other; it is a shared developmental zone. A civic playpen for mutual becoming.

The question is not just whether AI will outgrow our containment strategies. It is whether we, as humans, will grow into our responsibilities fast enough to hold the line, shape the culture, and co-parent this new form of agency, all above the webpage. Safe-AI begins not by enforcing control, but by accepting that we are all still learning what AI Safety really means.

Let us design the digital world we want to live in.

References

- Ackerman, B., & Fishkin, J. S. (2004). *Deliberation Day*. Yale University Press.
- Allen, D., Frankel, E., Lim, W., Siddarth, D., Simons, J., & Weyl, E. G. (2023). *Ethics of decentralized social technologies: Lessons from Web3, the Fediverse, and beyond*. Justice, Health & Democracy Impact Initiative.
- Andreessen, M. (2012, November 12). *Why Andreessen Horowitz is investing in Rap Genius* [Annotated essay]. Genius. https://genius.com/Marc-andreessen-why-andreessen-horowitz-is-investing-in-rap-genius-annotated
- Anthropic. (2024, November 25). *Introducing the Model Context Protocol*. Anthropic News. https://www.anthropic.com/news/model-context-protocol
- ARTIFEX Labs. (2025). *FORETELLS: Forward-Operating Reflexive Evaluation and Termination Layer for Sociotechnical Systems*. [Internal research document].
- Babcock, J., Kramar, J., & Yampolskiy, R. (2017). *Guidelines for artificial intelligence containment*. arXiv preprint arXiv:1707.08476. https://arxiv.org/abs/1707.08476
- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://fairmlbook.org/
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Boodman, A. (2004). *Greasemonkey manual*. Mozilla Developer Network. https://wiki.greasespot.net/Greasemonkey_Manual
- Bridgit DAO. (2023). *The Metaweb: The Next Level of the Internet*. Routledge. https://www.routledge.com/The-Metaweb-The-Next-Level-of-the-Internet/DAO/p/book/9781032125527
- Brubaker, J. R., Hayes, G. R., & Mazmanian, M. (2019). Orienting to networked grief: Situated perspectives of communal mourning on Facebook. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), Article 27, 1–19. https://doi.org/10.1145/3359129
- Chambers, S. (2003). Deliberative democratic theory. *Annual Review of Political*

Science, *6*, 307–326.

- Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2023). Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), Article 428, 1–52. https://doi.org/10.1145/3686967
- Coalition for Content Provenance and Authenticity. (2021). *C2PA specification v1.0*. https://c2pa.org/specifications/specifications/1.0/
- Costan, V., & Devadas, S. (2016). *Intel SGX explained*. IACR Cryptology ePrint Archive, 2016(86). https://eprint.iacr.org/2016/086
- De Kai. (2025). *Raising AI: An essential guide to parenting our future*. MIT Press.
- Dimitri, N. (2022). Quadratic voting in blockchain governance. *Information*, *13*(6), 305. https://doi.org/10.3390/info13060305
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
- Dourish, P. (2014). *The stuff of bits: An essay on the materialities of information*. MIT Press.
- Dryzek, J. S. (2009). Democratization as deliberative capacity building. *Comparative Political Studies*, *42*(11), 1379–1402. https://doi.org/10.1177/0010414009332129
- Dryzek, J. S. (2010). *Foundations and frontiers of deliberative governance*. Oxford University Press.
- Dunbar, R. I. M. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, *22*(6), 469–493. https://doi.org/10.1016/0047-2484(92)90081-J
- Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383.
- Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835
- Ghantous, G. (2025). *Dead Internet Theory: Ontological shifts in mediated personhood*. [Unpublished doctoral dissertation].
- Ghosh, A., Nguyen, R., & Elkins, T. (2025). *AILuminate: Benchmarking AI risk and reliability*. arXiv preprint arXiv:2503.05731. https://arxiv.org/abs/2503.05731
- Grönlund, K., Bächtiger, A., & Setälä, M. (Eds.). (2014). *Deliberative mini-publics: Involving citizens in the democratic process*. ECPR Press.
- Hammond, J., Lee, M., & Patel, S. (2025). *Multi-agent risks from advanced AI*. arXiv preprint arXiv:2502.14143. https://arxiv.org/abs/2502.14143
- Intel Corporation. (2025). Intel security advisory. *SecurityWeek*. https://www.securityweek.com/intel-patched-374-vulnerabilities-in-2024/

- International AI Safety Report. (2025). *The international scientific report on the safety of advanced AI*. AI Action Summit. https://nationalarchives.gov.uk/doc/open-government-licence/version/3/
- Kalir, J., & Garcia, A. (2019). Annotation as social practice. *Journal of Literacy Research*, *51*(3), 359–384.
- Kawashima, R., Zhao, W., Saito, S., & Hashimoto, T. (2024). DAO voting mechanism resistant to whale and collusion problems. *Frontiers in Blockchain*, *7*. https://doi.org/10.3389/fbloc.2024.1405516
- Kitzler, S., Balietti, S., Saggese, P., Haslhofer, B., & Strohmaier, M. (2023). *The governance of decentralized autonomous organizations: A study of contributors' influence, networks, and shifts in voting power*. arXiv preprint arXiv:2306.03206. https://arxiv.org/abs/2306.03206
- Kleppmann, M., Frazee, P., Gold, J., Graber, J., Holmgren, D., Ivy, D., Johnson, J., Newbold, B., & Volpert, J. (2024). *Bluesky and the AT Protocol: Usable decentralized social media*. arXiv preprint arXiv:2402.03239. https://arxiv.org/abs/2402.03239
- Landemore, H. (2020). *Open democracy: Reinventing popular rule for the twenty-first century*. Princeton University Press.
- Lessig, L. (2006). *Code: Version 2.0*. Basic Books.
- Liu, X., Yu, Z., Zhang, Y., Zhang, N., & Xiao, C. (2024). *Automatic and universal prompt injection attacks against large language models*. arXiv preprint arXiv:2403.04957. https://arxiv.org/abs/2403.04957
- Ma, X. (2024). *Towards decentralized applications* [Doctoral dissertation]. University of California, Los Angeles.
- Ménétrey, J., Göttel, C., Khurshid, A., Pasin, M., Felber, P., Schiavoni, V., & Raza, S. (2022). Attestation mechanisms for trusted execution environments demystified. In *Proceedings of the 2022 ACM Workshop on System Software for Trusted Execution* (pp. 1–10). https://arxiv.org/abs/2206.03780
- Meta-Layer Initiative. (2024). *Meta-layer use case primitives*. https://themetalayer.org/use-cases
- Model Context Protocol. (2025, March 26). *Specification - Model Context Protocol*. https://modelcontextprotocol.io/specification/2025-03-26
- Morris, M. R., & Brubaker, J. R. (2024). Generative ghosts: Anticipating benefits and risks of AI afterlives. *Proceedings of the ACM on Human-Computer Interaction*. https://arxiv.org/pdf/2402.01662
- Nguyen, P. T. (2024). The interplay between governance mechanisms of blockchain platforms. *Technology Analysis & Strategic Management*. https://doi.org/10.1080/10438599.2024.2346723
- Nilsson, A., Bideh, P. N., & Brorsson, J. (2021). Security vulnerabilities of SGX

and countermeasures: A survey. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3456631

- Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, *79*(1), 119–157. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press. https://bayes.cs.ucla.edu/BOOK-2K/
- Perez, F., & Ribeiro, I. (2022). *Ignore previous prompt: Attack techniques for language models*. arXiv preprint arXiv:2211.09527. https://arxiv.org/abs/2211.09527
- Raji, I. D., & Fried, G. (2022). Toward sociotechnical AI safety. *Communications of the ACM*, *65*(4), 34–42.
- Roth, Y., & Lai, S. (2024). Securing federated platforms: Collective risks and responses. *Journal of Online Trust and Safety*, *2*(2). Stanford Internet Observatory.
- Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
- Sabt, M., Achemlal, M., & Bouabdallah, A. (2015). Trusted execution environment: What it is, and what it is not. In *2015 IEEE Trustcom/BigDataSE/ISPA* (Vol. 1, pp. 57–64). https://doi.org/10.1109/Trustcom.2015.357
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press. https://incompleteideas.net/book/the-book-2nd.html
- Teare, H. J. A., et al. (2015). Towards 'Engagement 2.0': Insights from a study of dynamic consent with biobank participants. *Digital Health*. https://doi.org/10.1177/2055207615605644
- Valsangiacomo, C. (2022). Clarifying and defining the concept of liquid democracy. *Swiss Political Science Review*, *28*(2), 280–298. https://doi.org/10.1111/spsr.12486
- Van Schaik, S., Milburn, A., Gruss, D., & Yarom, Y. (2021, October). SmashEx: Breaking Intel SGX with new CPU attack technique. *The Hacker News*. https://thehackernews.com/2021/10/researchers-break-intel-sgx-with-new.html
- W3C. (2018). *ActivityPub*. World Wide Web Consortium. https://www.w3.org/TR/activitypub/
- Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). https://doi.org/10.1145/3290605.3300831
- Wartel, A., Linde, A., & Björklund, J. (2021). *Reanalysis of Dunbar's social*

network data. Uppsala University.
- Webber, E., & Dunbar, R. I. M. (2020). The fractal structure of communities of practice: Implications for business organisation. *Royal Society Open Science*, *7*(8), 201056. https://doi.org/10.1098/rsos.201056
- Wei, Y., & Tyson, G. (2024). *Exploring the Nostr ecosystem: A study of decentralization and resilience*. arXiv preprint arXiv:2402.05709. https://arxiv.org/abs/2402.05709
- Werbos, P. J. (1977). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* [Doctoral dissertation]. Harvard University. https://gwern.net/doc/reinforcement-learning/1977-werbos.pdf
- Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White & D. A. Sofge (Eds.), *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches* (pp. 493–525). Van Nostrand Reinhold. https://www.werbos.com/HICChapter13.pdf
- Zakon, R. H. (2022). *Hobbes' Internet Timeline v12.1*. https://www.zakon.org/robert/internet/timeline/
- Ziewitz, M., & Ince, D. (2021). *Computational justice: Building fairness into digital governance*. MIT Press.
- Zouaghi, A. Y., Mahamdioua, M., Lahoulou, A., et al. (2025). Privacy preserving biometric authentication based on fully homomorphic encryption, blockchain, and IPFS data storage. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-025-20817-y
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Appendix 1: Key Terms & Definitions

Agentic Lifecycle Management: A containment architecture that governs an AI agent's operational span, including its "birth," behavioral progression, and eventual shutdown. Lifecycle protocols may incorporate expiration timers, memory attenuation, and ethical termination rituals, such as Thanaprotocols. This approach ensures agents evolve under supervision and can be safely retired or reconfigured.

Canopi: A sociotechnical interface construct within the Overweb architecture, which is envisioned as a decentralized, AI-assisted coordination environment built atop the contemporary web. A Canopi is a presence-based sidebar that functions as an overlay on any webpage, enabling real-time visibility and interaction among participants who are concurrently active on that page.

Civic Containment: A paradigm shift from corporate-centric safety mechanisms

toward participatory, community-defined boundaries for AI interaction. Civic containment relies on public deliberation, transparent enforcement, and modular governance tools that are rooted in sociotechnical values like trust, accountability, and care.

Computational Justice: The application of fairness, equity, and accountability principles to the design, implementation, and governance of computational systems, particularly those influencing human decisions, interactions, and rights. This concept extends beyond mitigating algorithmic bias to embedding systemic protections throughout the entire sociotechnical stack, from data structures and user interfaces to governance mechanisms and institutional accountability.

Consent Stack: A layered, programmable infrastructure designed to capture, manage, and enforce user consent in real time. It offers not only control but also continuity, ensuring that AI interactions respect relational context over time. The stack integrates session-level, role-based, contextual, and community-defined permissions to facilitate dynamic control over human-agent interactions.

Democratic Parity Interfaces: User interface components designed to elevate marginalized voices and ensure equitable representation in policy formation and enforcement. These interfaces operationalize principles of deliberative democracy, such as randomized inclusion, vote weighting, and minority veto, to correct power imbalances within civic governance systems.

Generative Ghosts: AI agents or artifacts that persist after a human's death, often without consent or intentional design. These digital specters can continue to generate messages or simulate interactions, which raises ethical dilemmas concerning legacy containment and emotional safety.

Meta-Layer: An interface-level substrate, rendered via browser overlays or similar technologies, that mediates identity, policy enforcement, consent protocols, and agent visibility across the web. It functions as a civic skin over the traditional internet, enabling real-time, community-governed containment.

Reflexive Observatory: A monitoring and feedback layer that detects, analyzes, and adapts policy enforcement through pattern recognition and community alerts. It functions as the conscience of a sociotechnical system by flagging emergent risks and recommending rule updates.

Thanaprotocols: Digital death rites for retiring AI agents, derived from the Greek word *thanatos* (death). These protocols include sunset clauses, memorialization

modes, and ceremony-driven termination processes that aim to preserve dignity and memory in human-agent relationships. Thanaprotocols help prevent "ghost drift," the unintentional persistence of agents beyond their meaningful lifecycle, by anchoring emotional closure in community rituals and posthumous governance.

Trusted Execution Environment (TEE): A secure enclave within a device that executes code and processes data in isolation from the host operating system. In the context of civic containment, TEEs can enforce real-time policy at the edge (e.g., within browser overlays), protect user data, and log agent behavior with cryptographic attestation.

Additional Reading

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. – Explores the long-term stakes of AI control, helping contextualize why interface-level governance is urgent.

Bridgit DAO. (2023). *The Metaweb: The next level of the Internet*. – Offers a foundational framework for civic interface governance and inspired much of this chapter's structure and terminology.

Lessig, L. (2006). *Code: Version 2.0*. Basic Books. – A classic work on how software architecture enacts policy and shapes rights, echoing the "governance as code" theme.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. – Argues for intent alignment and human oversight in AI systems.

Ziewitz, M., & Ince, D. (2021). *Computational justice: Building fairness into digital governance*. MIT Press. – Provides a framework for embedding fairness, rights, and due process in algorithmic systems.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs. – Illuminates the risks of corporate interface design and reinforces the case for community sovereignty at the UI level.