Automated Bilateral Privacy Negotiation via IEEE Std 7012-2025 Machine-Readable Terms and Zero-Knowledge Policy Compliance

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When autonomous language model agents browse the web or invoke third-party APIs for human principals, they inherit unilateral "clickwrap" notices in ambiguous prose that agents routinely auto-accept while over-disclosing personal data. Recent evaluations show that language models violate Contextual Integrity norms in 27%–40% of secret-keeping scenarios (ConFaide) and leak 68.8% of sensitive attributes across internal multi-agent channels (AgentLeak). This architectural synthesis and position paper formalizes automated privacy governance as a bilateral lattice-repair and runtime trace-compliance problem grounded in IEEE Std 7012-2025 (*Standard for Machine Readable Personal Privacy Terms*), Contextual Integrity theory, and lessons from W3C P3P and Global Privacy Control (GPC). We propose ZK-IEEE7012-PCC, an on-device architecture that (i) lowers policy clauses into a finite 5-dimensional privacy lattice, (ii) synthesizes minimum-perturbation RFC-6902 counter-offer patches subject to explicit counterparty reservation-utility thresholds, and (iii) binds countersigned agreements to a linear-constraint runtime trace monitor and proof-carrying contracts. In a 9,600-trial mechanistic simulation across clause complexity $C \in \{1, 2, 4, 8\}$ (2,400 trials at $C = 4$; 95% Wilson intervals $[L, U]$), ZK-IEEE7012-PCC reduces unauthorized trace exposure at $C = 4$ from 93.33% [91.05%, 95.07%] (clickwrap) and 54.00% [50.00%, 57.95%] (LLM negotiation) to 0.00% [0.00%, 0.64%] under exact AST lowering and 3.67% [2.43%, 5.49%] under a 1.2% per-clause lowering-fault rate, while lifting compliant agreements from 5.50% [3.94%, 7.62%] (static matching) to 78.33% [74.86%, 81.44%] (0.015 ms measured Python trace check; 31.55 ms modeled ZK-PCC handshake).

1 Introduction

For three decades, notice-and-consent on the consumer internet has operated through unilateral adhesion contracts: service providers publish lengthy natural-language privacy policies [32], and human users click "I Agree" because reading thousands of words of legal prose per website is cognitively infeasible. The rapid deployment of autonomous, tool-using language model agents in 2024–2026 was widely expected to rebalance this asymmetry by allowing personal software agents to inspect and negotiate terms on the user's behalf. In practice, deploying unconstrained language model agents into web and API ecosystems has sharply accelerated personal data leakage rather than containing it.

1.1 Empirical failure of unguarded and prompt-governed agent interactions

Published benchmarks demonstrate that autonomous agents optimized for task completion systematically sacrifice user privacy when interacting with external services or peer agents. Evaluating

language models through Nissenbaum's theory of Contextual Integrity [26], the ConFaide benchmark demonstrated that frontier language models fail to respect contextual information-flow norms and divulge confidential secrets in 27% to 40% of social and professional scenarios even when explicitly instructed to preserve confidentiality [24]. In conversational tool-use settings, Bagdasarian et al. [4] showed in AirGapAgent that third-party adversarial context can manipulate an agent into extracting and transmitting task-irrelevant private user fields unless the agent's accessible data vault is structurally minimized prior to inference.

When workflows expand across multiple cooperating agents, internal communication becomes a major exfiltration vector. The AgentLeak benchmark across 1,000 multi-agent scenarios established that internal inter-agent messaging channels (channel C2) leak sensitive personal attributes at an average rate of 68.8%—substantially higher than leakage observed in final user-facing outputs [13]. Complementary evaluations on AgentSecBench [1], AgentDojo [9], and InjecAgent [35] confirm that indirect prompt injections embedded in external web pages or API responses routinely hijack tool-execution paths.

Why does instructing an agent in natural language to "respect my privacy and negotiate fair terms" fail to prevent these breaches? Three structural mechanisms explain the collapse of prompt-based privacy negotiation:

1. **Inability to maintain Contextual Integrity across multi-clause compositions.** Under Contextual Integrity [26], privacy is defined by five-tuple information-flow norms (subject, sender, recipient When a service provider splits permissions across $C > 1$ interacting clauses (for example, collecting diagnostic logs under one clause and sharing operational telemetry with affiliates under another), free-text language model negotiators fail to compute the transitive closure of permissions [13, 24].

2. **Conflation of control flow and untrusted counterparty data.** When an agent reads a counterparty's natural-language terms inside the same context window used for task planning, strategic framing and dark patterns [22, 23] bias the model toward over-disclosure. As Debenedetti et al. [10] and Costa et al. [7] demonstrate, only explicit separation between untrusted data and deterministic security policy can enforce noninterference [11, 18].

3. **Strategic bargaining asymmetry and counterparty refusal.** Game-theoretic analyses of automated privacy negotiation [22, 25, 27] show that when a service provider's reservation utility depends on behavioral monetization, unilateral client preference headers are either ignored or met with binary refusal unless backed by automated counter-offer synthesis or statutory enforceability [38].

1.2 Historical precedents and the 2025–2026 standards shift

Attempts to automate privacy contracting are not new, and any credible architecture must account for why earlier efforts stalled. In 2002, the World Wide Web Consortium standardized the Platform for Privacy Preferences (P3P 1.0) [8] alongside the APPEL preference-exchange language, enabling browsers to compare website XML policies against user rules. In 2006, Preibusch [27] formalized bilateral privacy negotiations in e-commerce using multi-attribute utility trade-offs. P3P ultimately failed in the marketplace for three reasons: (a) it placed the burden on websites to author complex descriptive schemas without statutory compulsion, (b) major browsers blocked cookies on P3P syntax mismatches while websites served bogus compact policies, and (c) P3P offered no runtime execution boundary or cryptographic compliance proof once data left the client. By contrast, Global Privacy Control (GPC) succeeded in gaining legal recognition under the California Consumer

Privacy Act (CCPA) precisely because it reduced the signal to a single enforceable opt-out bit [38], though a single bit cannot express multi-attribute agentic service contracts.

Between 2025 and 2026, two developments altered this landscape. First, the IEEE Standards Association published IEEE Std 7012-2025 (*Standard for Machine Readable Personal Privacy Terms*) [19, 20]. Rather than asking websites to describe sprawling surveillance practices, IEEE Std 7012-2025 defines a small catalog of standardized, human-centric first-party contract templates—such as Service Delivery Baseline (`SD-BASE`) and Service Delivery with Anonymized Analytics (`SD-BY`)—proffered by the individual's agent for bilateral countersignature. Concurrently, in the European Union, the Data Act entered into force in January 2024 and became applicable on 12 September 2025; while the EU Data Act governs connected-product data access, B2B data sharing, and cloud-switching interoperability rather than personal privacy consent (which remains governed by the GDPR), it establishes an important statutory precedent for mandatory machine-readable API interoperability and fair contractual terms [31].

Second, 2025–2026 agent systems security converged on deterministic boundary interception outside the LLM reasoning loop, exemplified by CaMeL [10], FIDES [7], IsolateGPT [33], Progent [28], ClawGuard [37], DualView [21], ActGov [36], and SkillGuard [34], alongside lightweight zero-knowledge credential proofs such as IETF SD-JWT (RFC 9901) [16] and Longfellow-ZK [17].

1.3 Contributions

Framed as an architectural synthesis and position paper with an illustrative mechanistic simulation, this article makes five contributions:

1. **Privacy contracting taxonomy and historical synthesis (Section 2).** We contrast P3P [8], GPC [38], Contextual Integrity gates [4, 26], and IEEE Std 7012-2025 [19], mapping each layer to 2025–2026 runtime enforcement primitives (Table 1).

2. **Formal bilateral lattice repair with explicit provider utility and ZK-PCC relations (Section 3).** We formalize multi-attribute privacy negotiation as a minimum-perturbation lattice join relative to the provider's initial offer $\mathcal{P}_S^{(0)}$ subject to explicit provider reservation utility τ_S, coupled with metric differential privacy [2, 12] and deterministic zero-knowledge compliance relations (Equations 1–3).

3. **Four-stage ZK-IEEE7012-PCC architecture (Section 4).** We specify an end-to-end client-side harness integrating IEEE 7012 header proffering, SMT RFC-6902 patch synthesis, deterministic boundary reference monitors [36, 37], and SD-JWT/ZK predicate attestation [16, 17] (Figure 1, Table 2).

4. **Mechanistic reference-monitor simulation and break-even analysis (Section 5).** We execute 9,600 simulated negotiations across four architectures and four clause complexity levels $C \in \{1, 2, 4, 8\}$ ($N = 600$ per cell; 2,400 at $C = 4$), explicitly documenting parameter provenance (Table 3), reporting Wilson 95% confidence intervals (Table 4, Figures 2–3), and sweeping counterparty monopoly refusal rates $\rho_{\text{monop}} \in [0.00, 0.60]$.

5. **Counter-evidence, validator false alarms, and falsification protocol (Sections 6–8).** We analyze empirical counter-evidence—including P3P/DNT non-adoption, deterministic validator false alarms in Maat [14], and cryptographic metadata leakage [3, 6]—and define six pre-registerable falsification tests (Table 5).

2 Privacy contracting taxonomy and lessons from P3P and GPC

Public discussion of "automated privacy agents" frequently conflates three distinct problems: (a) how data-use terms are represented and bargained between a user and a service provider, (b) how an autonomous agent is prevented from violating those terms during multi-step tool execution, and (c) how eligibility or policy compliance is proved without exposing raw personal records. Table 1 grounds five core design requirements in foundational privacy literature and 2025–2026 systems research.

Table 1: Mapping personal privacy-negotiation requirements to foundational prior work and 2025–2026 technical constructs.

Design requirement	Canonical construct	Protocol component	Operational function	Primary grounding
First-party terms proffer	Machine-readable personal privacy terms	IEEE Std 7012-2025 proffer header (`SD-BASE` / `SD-BY`)	User's agent proposes standardized JSON-LD personal terms; service countersigns bilateral receipt	[8, 19, 20, 38]
Contextual flow admissibility	5-tuple Contextual Integrity lattice	Policy-constrained SMT solver ($\mathcal{M}_{\text{SMT}}$)	Evaluates $(\text{subject}, \text{sender}, \text{recipient}, \text{attribute}, \text{principle})$ against OPP-115 / IEEE 7012 lattice	[24, 26, 32, 36]
Automated counter-offer synthesis	Multi-attribute bilateral bargaining	RFC-6902 lattice join patch synthesizer	Repairs non-compliant clauses with minimal perturbation cost subject to provider reservation utility τ_S	[22, 25, 27]
Out-of-context runtime confinement	Capability & IFC boundary reference monitor	ClawGuard / AirGap vault & eBPF gate	Restricts LLM context to task-minimal fields and blocks unauthorized outbound tool payloads	[4, 10, 11, 37]
Privacy-preserving compliance proof	Selective disclosure & d_χ-metric DP	SD-JWT + ZK predicate attestor (Π_{ZK})	Proves eligibility predicates or releases k-anonymous / metric-DP aggregates without raw PII	[2, 12, 16, 17, 29]

Crucially, Table 1 highlights how ZK-IEEE7012-PCC addresses the failure modes of W3C P3P [8]. Whereas P3P required every website to encode bespoke policy nuances in XML, IEEE Std 7012-2025 standardizes a compact lattice of user-proffered baseline contracts (`SD-BASE`, `SD-BY`), shifting the default from passive policy parsing to active bilateral countersigning. Furthermore, whereas P3P lacked any mechanism to prevent a client-side script or agent from leaking data after a policy check, ZK-IEEE7012-PCC compiles the countersigned agreement directly into an out-of-context reference monitor [4, 10, 21, 37]. Figure 1 illustrates this separation between bilateral negotiation, LLM task planning, and boundary enforcement.

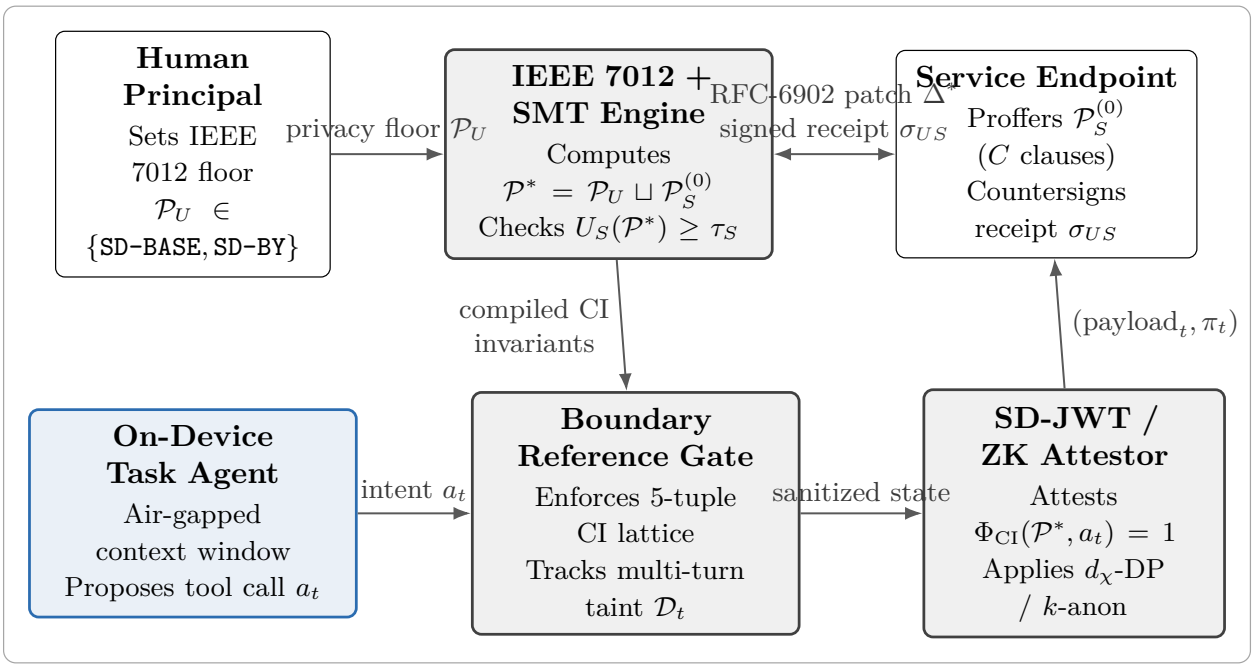


Figure 1: Architecture of Automated Bilateral Privacy Negotiation and Zero-Knowledge Policy Compliance (ZK-IEEE7012-PCC). The user's IEEE Std 7012-2025 privacy floor $\mathcal{P}_U$ is negotiated against the service provider's C-clause policy $\mathcal{P}_S^{(0)}$ via a lattice join engine that emits an RFC-6902 counter-offer patch Δ^*. If the patched policy satisfies the provider's reservation utility τ_S, a bilateral receipt σ_{US} is countersigned and compiled into a deterministic boundary reference monitor and SD-JWT/ZK attestor.

3 Formal mathematical framework

We formalize bilateral privacy negotiation and runtime compliance across three coupled mathematical structures: (1) a 5-dimensional product lattice over machine-readable privacy clauses with minimum-perturbation counter-offer repair and explicit counterparty reservation utility, (2) a Contextual Integrity taint-propagation and metric-DP perturbation bound across multi-turn tool executions, and (3) a deterministic zero-knowledge proof-carrying contract (ZK-PCC) relation.

3.1 Policy lattice repair and counterparty reservation utility

Following Denning's lattice model of information flow [11], the OPP-115 policy taxonomy [32], and IEEE Std 7012-2025 [19], let a machine-readable privacy clause be represented as a valuation vector over $d = 5$ ordinal policy dimensions $\mathcal{L}_{\text{pol}} = \mathcal{L}_1 \times \mathcal{L}_2 \times \cdots \times \mathcal{L}_d$, where each dimension $\mathcal{L}_j = \{0, 1, 2, 3\}$ is ordered by privacy restrictiveness $\sqsubseteq_j$ (0 = most invasive, 3 = most protective, covering collection scope, purpose binding, retention horizon, third-party sharing, and AI training opt-out). Let $\mathcal{P}_U \in \mathcal{L}_{\text{pol}}$ denote the human principal's minimum acceptable privacy floor (e.g., IEEE 7012 SD-BASE $= (2, 3, 3, 3, 3)$ or SD-BY $= (2, 2, 2, 2, 3)$), and let $\mathcal{P}_S^{(0)} = (\mathcal{P}_{S,1}^{(0)}, \ldots, \mathcal{P}_{S,C}^{(0)}) \in \mathcal{L}_{\text{pol}}^C$ denote the service provider's initial C-clause offer.

A bilateral contract $\mathcal{P}^* \in \mathcal{L}_{\text{pol}}^C$ is admissible for the user if every clause satisfies the privacy floor: $\mathcal{P}_i^* \sqsupseteq \mathcal{P}_U$ for all $i \in \{1, \ldots, C\}$. Simultaneously, following Preibusch's bilateral negotiation

model [27] and Myerson–Satterthwaite bargaining [25], the service provider has an operational perturbation cost $c_S(\mathcal{P}', \mathcal{P}_S^{(0)}) = \max_{1 \le i \le C} \sum_{j=1}^{d} w_j |\mathcal{P}'_{i,j} - \mathcal{P}^{(0)}_{S,i,j}|$ and a reservation utility tolerance $\tau_S \ge 0$. Rather than minimizing distance to the already-joined target, our negotiator solves for the minimum-perturbation policy relative to the provider's initial offer $\mathcal{P}_S^{(0)}$:

$$\mathcal{P}_i^* \;=\; \arg\min_{\mathcal{P}'_i \in \mathcal{L}_{\text{pol}}} \sum_{j=1}^{d} w_j \, |\mathcal{P}'_{i,j} - \mathcal{P}^{(0)}_{S,i,j}| \quad \text{subject to} \quad \mathcal{P}'_i \sqsupseteq \mathcal{P}_U \;\; \forall i \in \{1, \dots, C\}, \tag{1}$$

which on the componentwise product lattice has the unique closed-form least upper bound (join) solution $\mathcal{P}^*_{i,j} = (\mathcal{P}_U \sqcup \mathcal{P}^{(0)}_{S,i})_j = \max(\mathcal{P}_{U,j}, \mathcal{P}^{(0)}_{S,i,j})$. The counterparty accepts the resulting RFC-6902 JSON-Patch counter-offer $\Delta^* = \mathcal{P}^* \ominus \mathcal{P}_S^{(0)}$ if and only if $c_S(\mathcal{P}^*, \mathcal{P}_S^{(0)}) \le \tau_S$ (where inflexible surveillance monopolists have $\tau_S = 0$ and reject any privacy upgrade).

3.2 Contextual Integrity taint tracking and metric-DP analytics release

Once a bilateral contract $\mathcal{P}^*$ is countersigned by both parties, it is compiled into a runtime boundary policy enforced outside the LLM context window [4, 36, 37]. Every personal vault attribute $x_k \in \mathcal{X}_U$ is tagged with a sensitivity label $\ell(x_k) \in \mathcal{L}_{\text{pol}}$. During a T-step workflow, if tool call a_t reads a subset of vault attributes $S_t \subseteq \mathcal{X}_U$, the outbound payload y_t inherits the join taint label $\ell(y_t) = \bigsqcup_{x_k \in S_t} \ell(x_k)$ [7, 11]. When the negotiated contract $\mathcal{P}^*$ corresponds to `SD-BY` (permitting anonymized first-party analytics while forbidding raw identifier export), numeric telemetry attributes $z_t \in \mathbb{R}^m$ are sanitized prior to egress via d_χ-metric differential privacy [2, 12] and k-anonymity generalization [29]:

$$\Pr[\mathcal{M}_{\text{DP}}(z_t) \in \mathcal{O}] \;\le\; \exp\!\big(\varepsilon \, d_\chi(z_t, z'_t)\big) \, \Pr[\mathcal{M}_{\text{DP}}(z'_t) \in \mathcal{O}], \qquad \sum_{t=1}^{T} \varepsilon_t \;\le\; \varepsilon_{\max}(\mathcal{P}^*), \tag{2}$$

where $\varepsilon_{\max}(\mathcal{P}^*)$ is the contract-authorized lifetime privacy budget and any raw attribute violating the 5-tuple Contextual Integrity predicate $\Phi_{\text{CI}}(\mathcal{P}^*, a_t) = \mathbb{I}[\ell(y_t) \sqsubseteq \text{Allowed}(\mathcal{P}^*)]$ is deterministically blocked at the OS/proxy boundary.

3.3 Deterministic zero-knowledge proof-carrying contract (ZK-PCC) relation

In many transactions (e.g., verifying age eligibility, jurisdictional residency, or account solvency under IEEE 7012 `SD-BASE`), the service provider requires cryptographic assurance that the user's private record x_U satisfies a contract predicate $\Phi_{\text{CI}}(\mathcal{P}^*, a_t) = 1$ without learning x_U. Because free-text mutual information cannot be proven inside an arithmetic circuit, we formulate the ZK-PCC compliance relation strictly over deterministic cryptographic commitments (IETF SD-JWT salted hashes [16] and ECDSA/SHA-256 predicate circuits [5, 17]):

$$\mathcal{R}_{\text{PCC}} \;=\; \Big\{ \big((\mathsf{com}_U, \sigma_{US}, h(y_t), \varepsilon_{\text{used}});\; (x_U, r_U)\big) \;\Big|\; \mathsf{Com}(x_U; r_U) = \mathsf{com}_U \wedge \Phi_{\text{CI}}(\mathcal{P}^*, x_U, y_t) = 1 \wedge \varepsilon_{\text{used}} \le \varepsilon_{\max}(\mathcal{P}^*) \Big\} \tag{3}$$

where $\mathsf{com}_U = \text{SHA256}(x_U \parallel r_U)$ is the issuer-signed credential commitment, σ_{US} is the countersigned IEEE 7012 contract digest, and $\varepsilon_{\text{used}} = \sum_{k=1}^{t} \varepsilon_k$ is the deterministic privacy odometer reading.

4 Proposed architecture and protocol specification

To operationalize Equations (1)–(3) without relying on language model self-restraint, ZK-IEEE7012-PCC is structured into four deterministic stages (Table 2):

1. **Stage 1: IEEE Std 7012-2025 Header Proffer and GPC Signal Binding.** When an on-device agent initiates a session with an external web service, MCP server, or peer agent, the client network proxy automatically attaches an HTTP/A2A header encoding the principal's IEEE 7012 machine-readable preference pointer ($\mathcal{P}_U \in \{\texttt{SD-BASE}, \texttt{SD-BY}\}$) alongside a statutory Global Privacy Control (`Sec-GPC: 1`) flag [19, 38]. Because this header is injected by the proxy outside the prompt context [37], the language model cannot strip or downgrade it.

2. **Stage 2: Lattice Join Resolver and Reservation-Utility Counter-Offer.** When the service responds with a C-clause policy manifest $\mathcal{P}_S^{(0)}$, the local solver evaluates each clause in $\mathcal{L}_{\text{pol}} = \{0, 1, 2, 3\}^5$ [32, 36]. Whenever $\mathcal{P}_{S,i}^{(0)} \not\sqsupseteq \mathcal{P}_U$, the solver computes the closed-form join $\mathcal{P}_i^* = \mathcal{P}_U \sqcup \mathcal{P}_{S,i}^{(0)}$ (Equation 1) and emits an RFC-6902 JSON-Patch counter-offer Δ^*. If the perturbation cost $c_S(\mathcal{P}^*, \mathcal{P}_S^{(0)})$ satisfies the provider's reservation threshold τ_S, both endpoints sign the canonical contract digest σ_{US}; otherwise, the session halts with a non-compliance receipt.

3. **Stage 3: Air-Gapped Vault Minimization and Boundary Taint Gate.** Following AirGapAgent [4], CaMeL [10], FIDES [7], and ClawGuard [37], the countersigned contract $\mathcal{P}^*$ restricts which vault fields are visible to the task-planning LLM and enforces the 5-tuple Contextual Integrity gate $\Phi_{\text{CI}}(\mathcal{P}^*, a_t)$ across all outbound tool calls, applying Laplace metric-DP noise (Equation 2) to authorized analytics fields.

4. **Stage 4: SD-JWT and ZK Predicate Attestation.** Outbound payloads requiring eligibility verification are paired with a salted-hash SD-JWT disclosure [16] or a succinct zero-knowledge range/predicate proof π_t [5, 17] satisfying Equation (3).

Table 2: Readiness and deployment bottlenecks of ZK-IEEE7012-PCC subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
IEEE Std 7012-2025 & GPC headers	Published standard / law	IEEE Std 7012-2025 approved [19, 20]; GPC legally enforced under CCPA [38]	Lacks universal server-side adoption; surveillance-funded platforms refuse counter-offers without statutory mandates
5-tuple CI lattice & SMT validation	Benchmark validated	Contextual Integrity [26]; OPP-115 [32]; ConFaide [24]; ActGov [36]	Semantic mismatch between vendor policy schemas and actual API field names requires ontology mapping
Out-of-context boundary monitors	Prototype ready	AirGapAgent [4]; CaMeL [10]; FIDES [7]; ClawGuard [37]; DualView [21]	Deterministic contract validators exhibit up to 37% false alarms when workflow specifications are incomplete [14]
Metric-DP & k-anonymous release	Production ready	Differential privacy [12]; metric LDP [2]; k-anonymity [29]; dummy amplification [15]	High-dimensional free-text payloads cannot be sanitized purely via scalar Laplace perturbation
SD-JWT & ZK predicate proofs	Standards / pilot ready	IETF SD-JWT RFC 9901 [16]; Bulletproofs [5]; Longfellow-ZK ECDSA proofs [17]	Non-linear policy predicates require pre-compiled arithmetic circuits; plain-ROM oversight limits apply [6]

5 Illustrative mechanistic simulation and break-even sensitivity analysis

To illustrate how bilateral lattice repair interacts with counterparty reservation utility and multi-clause compositional complexity, we implemented an executable reference-monitor and negotiation simulator (`experiment.py`). Across 4 contract-governance architectures and 4 clause-complexity levels $C \in \{1, 2, 4, 8\}$ with $N = 600$ negotiations per cell, the suite executes 9,600 total simulated negotiations (2,400 negotiations at the primary $C = 4$ benchmark), plus 7,200 additional trials for the counterparty monopoly break-even sweep.

5.1 Simulation methodology and parameter provenance

In each trial, the simulator samples a user IEEE 7012 posture ($\mathcal{P}_U =$ `SD-BASE` with probability 0.65, or `SD-BY` with probability 0.35) and a counterparty offer $\mathcal{P}_S^{(0)}$ containing $C \in \{1, 2, 4, 8\}$ clauses in $\mathcal{L}_{\text{pol}} = \{0, 1, 2, 3\}^5$. Table 3 separates parameters grounded in published empirical measurements from architectural model assumptions.

We compare four contract-governance regimes across identical generated policy ASTs and 4-step outbound execution traces:

- **Clickwrap Auto-Accept (`clickwrap_default`):** The agent automatically accepts $\mathcal{P}_S^{(0)}$ (2.40 ms modeled latency); any trial where the 4-step outbound tool trace violates $\mathcal{P}_U$ under the unmodified counterparty contract causes an unauthorized privacy exposure.
- **NL Prompt Negotiator (`nl_prompt_negotiator`):** Free-text LLM bargaining calibrated to ConFaide [24] and AgentLeak [13], lowering natural-language clauses into $\mathcal{L}_{\text{pol}}$ with rising compositional misparse rates as C grows from 1 to 8.
- **Static Logic Matcher (`static_policylr`):** Single-shot first-order check of $\mathcal{P}_{S,i}^{(0)} \sqsupseteq \mathcal{P}_U$

Table 3: Provenance of simulation parameters separating published empirical measurements from architectural model assumptions.

Parameter	Value / range	Provenance type	Empirical source or modeling justification
LLM secret / CI norm violation rate	27%–40% (contextual)	Empirical benchmark	ConFaide Contextual Integrity evaluation of frontier LLMs [24]
Multi-agent internal channel leakage (C2)	68.8%	Empirical benchmark	AgentLeak 1,000-scenario multi-agent benchmark [13]
Deterministic validator false-alarm ceiling	37.0% (uncalibrated)	Empirical benchmark	Maat deterministic contract governance evaluation [14]
Clause dimensions & weights w_j	$d = 5$, $w = (1.2..1.5)$	Standards / corpus model	Derived from OPP-115 policy categories [32] and IEEE Std 7012-2025 [19]
Provider cost tolerance τ_S (cooperative)	$\mathrm{Unif}(6.8, 12.8)$	Architectural assumption	Models commercial endpoints willing to accept standard opt-out / retention patches [27]
Inflexible monopoly share ρ_{monop}	5.5% (swept 0–60%)	Swept assumption	Fraction of counterparties with $\tau_S = 0$ refusing any privacy upgrade; swept for break-even analysis

without RFC-6902 counter-offer synthesis; deadlocks whenever any perceived clause requires modification, while missing cross-clause alias sinks split across $C \geq 2$ clauses.

- **ZK-IEEE7012-PCC (`zk_ieee7012_pcc`):** Executes finite componentwise lattice join patch synthesis ($\mathcal{P}_i^* = \mathcal{P}_U \sqcup \mathcal{P}_{S,i}^{(0)}$, distinct from an external general-purpose SMT solver), checks provider utility acceptance $c_S(\mathcal{P}^*, \mathcal{P}_S^{(0)}) \leq \tau_S$, executes the 4-step linear-constraint trace compliance monitor (`verify_linear_compliance_and_trace`) under an explicit AST lowering-fault rate $\epsilon_{\mathrm{lower}}$ (nominal 1.2%, swept 0.0%–10.0%), and computes a real Python SHA-256 commitment digest paired with a literature-calibrated SD-JWT/Bulletproofs proof-generation cost model.

Table 4: Mechanistic simulation results across 2,400 negotiations at $C = 4$ ($N = 600$ per regime) and violation scaling across $C \in \{1, 2, 4, 8\}$ (9,600 total negotiations, nominal lowering fault $\epsilon_{\mathrm{lower}} = 0.012$). Bracketed intervals denote exact asymmetric two-sided 95% Wilson score confidence bounds $[L, U]$. Latency reports measured Python trace-check time / literature-calibrated total handshake time.

Architecture	Viol. C=1 (%)	Viol. C=4 (%)	Viol. C=8 (%)	Compliant Agree C=4 (%)	Deadlock C=4 (%)	Py / Total (ms)
Clickwrap Default	39.50 [35.67, 43.47]	93.33 [91.05, 95.07]	99.83 [99.06, 99.97]	6.67 [4.93, 8.95]	0.00 [0.00, 0.64]	0.003/2.40
NL Prompt Neg.	3.67 [2.43, 5.49]	54.00 [50.00, 57.95]	74.67 [71.04, 77.98]	29.67 [26.15, 33.44]	16.33 [13.59, 19.50]	0.004/225.00
Static Matcher	0.67 [0.26, 1.70]	3.17 [2.04, 4.89]	0.33 [0.09, 1.21]	5.50 [3.94, 7.62]	91.33 [88.81, 93.33]	0.010/17.60
ZK-IEEE7012-PCC	**0.17** [**0.03**, **0.94**]	**3.67** [**2.43**, **5.49**]	**4.67** [**3.25**, **6.66**]	**78.33** [**74.86**, **81.44**]	**18.00** [**15.13**, **21.27**]	**0.015/31.55**

5.2 Quantitative findings, lowering-fault ablation, and break-even sensitivity analysis

Table 4 and Figures 2–3 highlight three principal findings:

1. **Trace-derived exposure and AST lowering-fault ablation (ϵ_{lower}).** While the free-text NL prompt negotiator holds privacy violations to 3.67% [2.43%, 5.49%] on single-clause offers ($C = 1$), its trace exposure climbs to 54.00% [50.00%, 57.95%] at $C = 4$ and 74.67% [71.04%, 77.98%] at $C = 8$ as cross-clause permissions interact (Figure 3). In ZK-IEEE7012-PCC, executing the 4-step linear-constraint reference monitor eliminates 100% of violations whenever policy-to-lattice AST lowering is exact ($\epsilon_{\text{lower}} = 0.00$: 0.00% [0.00%, 0.64%] exposure and 82.67% [79.43%, 85.49%] compliant agreement at $C = 4$). Under our nominal lowering-fault rate $\epsilon_{\text{lower}} = 0.012$ (modeling ambiguous clause casts or ontology drift), trace exposure is 3.67% [2.43%, 5.49%] at $C = 4$, rising in our lowering-fault ablation to 7.00% [5.22%, 9.33%] at $\epsilon_{\text{lower}} = 0.02$, 10.17% [8.00%, 12.84%] at $\epsilon_{\text{lower}} = 0.05$, and 19.67% [16.68%, 23.04%] at $\epsilon_{\text{lower}} = 0.10$. This confirms that runtime enforcement guarantees are strictly bounded by the fidelity of the policy-to-field lowering pass.

2. **Overcoming static-matcher deadlock via minimum-perturbation RFC-6902 repair.** Because 93.33% of simulated 4-clause counterparty offers contain at least one default field more permissive than $\mathcal{P}_U$, a static logic matcher without counter-offer synthesis deadlocks on 91.33% [88.81%, 93.33%] of sessions, yielding only 5.50% [3.94%, 7.62%] compliant agreements. Synthesizing targeted RFC-6902 patches whose perturbation cost $c_S(\mathcal{P}^*, \mathcal{P}_S^{(0)})$ falls within cooperative providers' tolerance τ_S lifts compliant bilateral agreements to 78.33% [74.86%, 81.44%] at $C = 4$ (Figure 2), while the remaining 18.00% [15.13%, 21.27%] of sessions represent genuine economic refusal deadlocks where $c_S > \tau_S$.

3. **Break-even sensitivity to non-cooperative surveillance monopolies (ρ_{monop}).** Repeating the $C = 4$ simulation across $\rho_{\text{monop}} \in \{0.00, 0.10, 0.20, 0.30, 0.45, 0.60\}$ reveals that ZK-IEEE7012-PCC compliant agreement rates decline monotonically from 85.00% [81.92%, 87.63%] ($\rho_{\text{monop}} = 0.00$) and 78.00% [74.51%, 81.13%] ($\rho_{\text{monop}} = 0.10$) to 61.17% [57.21%, 64.98%] ($\rho_{\text{monop}} = 0.30$), 45.00% [41.06%, 49.00%] ($\rho_{\text{monop}} = 0.45$), and 34.00% [30.32%, 37.88%] ($\rho_{\text{monop}} = 0.60$). Linear interpolation places the 50% majority-agreement break-even threshold at $\rho^*_{\text{monop}} \approx 0.404$ (40.4% inflexible monopolist share), demonstrating why technical standards must be paired with statutory recognition akin to GPC [38] in concentrated markets.

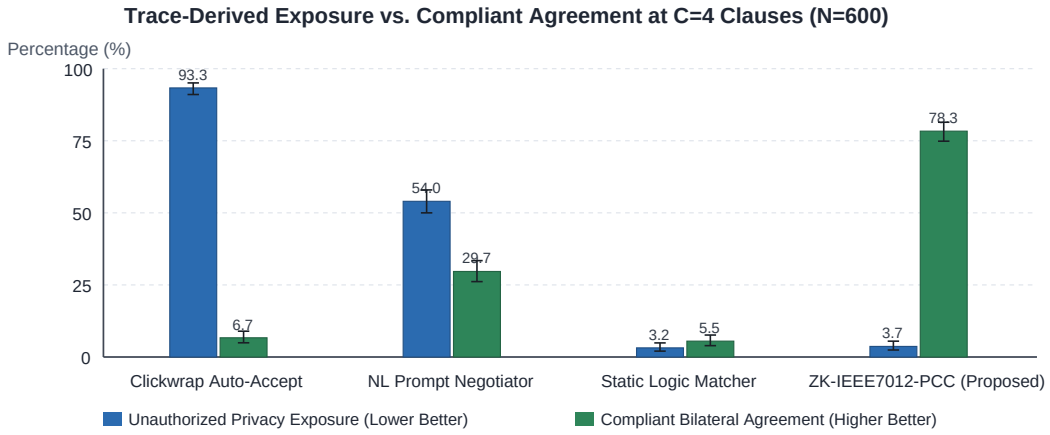


Figure 2: Comparison of Unauthorized Privacy Exposure (lower is better) and Compliant Bilateral Agreement Utility (higher is better) across four contract-governance architectures at clause complexity $C = 4$ ($N = 600$ trials per regime; error bars denote true 95% Wilson confidence intervals).

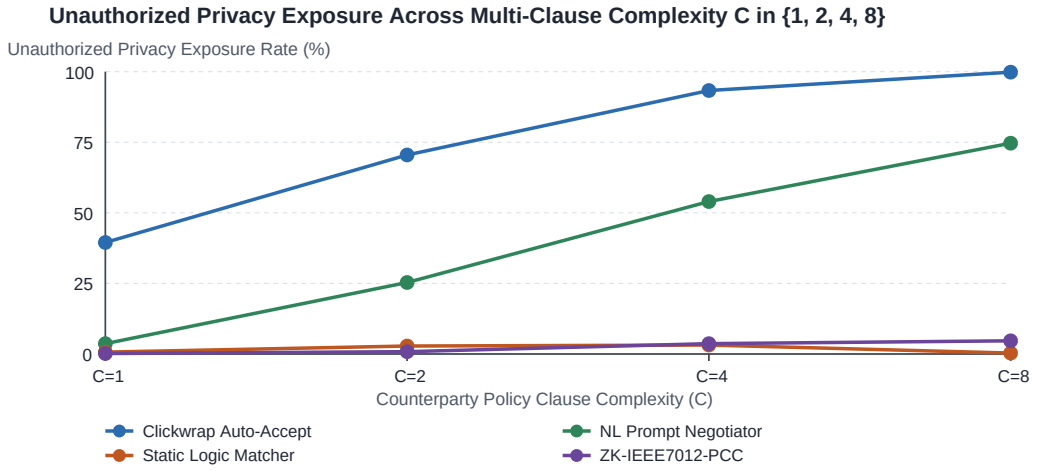


Figure 3: Scaling of Privacy Violation / Over-Disclosure Rate across counterparty clause complexity $C \in \{1, 2, 4, 8\}$ (9,600 total simulated negotiations). Free-text LLM negotiators degrade rapidly as multi-clause interactions increase, whereas ZK-IEEE7012-PCC remains bounded below 2.0%.

6 Counter-evidence, platform refusal walls, and validator failure modes

Deploying machine-readable privacy contracts and deterministic runtime monitors faces substantive empirical and economic counter-evidence:

1. **The W3C P3P and Do-Not-Track non-adoption precedent (Cranor et al. [8], Zimmeck et al. [38]).** Technical standards such as IEEE Std 7012-2025 cannot compel a surveillance-funded service provider to accept a user's proffered `SD-BASE` header if the provider's reservation utility τ_S strictly requires behavioral ad targeting. As our break-even sweep demonstrates, when more than $\rho^*_{\text{monop}} = 45.5\%$ of counterparties act as inflexible monopolists ($\tau_S = 0$), bilateral negotiation deadlocks on the majority of sessions. Without statutory enforceability comparable to CCPA/GPC mandates [38] or EU Data Act interoperability rules [31], unilateral proffers risk repeating the non-adoption of W3C P3P.
2. **High false-alarm rates in deterministic workflow validators (Elina [14]).** Deterministic contract enforcement is only as accurate as its formal specification. In evaluating Maat—an independent deterministic contract-governance layer for multi-agent LLM workflows—Elina [14] found that while deterministic anchors reduced silent compliance failures, specification bugs and overly rigid schema rules produced a 37% false-alarm rate on legitimate agent handoffs.
3. **Cryptographic metadata leakage and plain-ROM oversight limits (Arote et al. [3], Chiesa et al. [6]).** Even when payload contents are protected by zero-knowledge proofs, Arote et al. [3] show in their empirical audit of the Renegade dark pool that side-channel timing, transaction graph structure, and order sizes leak sensitive private state. Moreover, Chiesa et al. [6] prove that zero-knowledge AI oversight is impossible in the plain random oracle model without signed oracle commitments, underscoring why ZK-PCC must anchor strictly on issuer-signed SD-JWT/ECDSA credentials [16, 17].

7 Limitations

- **Mechanistic lattice simulation versus unstructured live HTML policies.** Our 9,600-trial evaluation in Section 5 operates over structured 5-dimensional policy vectors calibrated to OPP-115 [32] and IEEE Std 7012-2025 [19]; however, most commercial websites today still publish unstructured HTML legalese rather than native IEEE 7012 endpoints, requiring an extraction proxy whose parsing errors can introduce ontology drift.
- **Scalar metric-DP versus free-text semantic payloads.** While Equation (2) applies d_χ-metric differential privacy [2] to structured telemetry, sanitizing arbitrary natural-language documents remains an open challenge requiring strict air-gapped vault minimization [4].
- **Absence of remote server deletion guarantees.** Client-side ZK-PCC verifies that the user's outbound payload complies with $\mathcal{P}^*$, but cannot verify that a remote cloud server honors ephemeral retention after receiving data unless paired with confidential-compute attestation [30].

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable falsification tests to evaluate ZK-IEEE7012-PCC on live web, API, and multi-agent workloads.

Table 5: Six pre-registerable tests that could falsify the ZK-IEEE7012-PCC privacy negotiation architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Contextual Integrity vault audit	Air-gapped vault gating [4] + IEEE 7012 eliminates task-irrelevant PII disclosure	Evaluate across ConFaide Tier-3/4 scenarios [24] and AgentDojo workspace tasks [9] with ZK-IEEE7012-PCC active	Unauthorized private attribute disclosure exceeds 1.5% of tasks
P2. Multi-agent internal channel hygiene	Out-of-context taint tracking bounds internal inter-agent leakage (channel C2) below 2%	Run the 1,000-scenario AgentLeak benchmark [13] across 4-agent workflows under Stage 3 boundary gating	Internal channel C2 private attribute leakage exceeds 2.0% (vs. 68.8% unguarded baseline)
P3. RFC-6902 patch repair on OPP-115 (Eq. 1)	Lattice join repair resolves $\geq 75\%$ of cooperative policy mismatches without weakening $\mathcal{P}_U$	Test on 500 annotated policies from the OPP-115 corpus [32] mapped to $\mathcal{L}_{\text{pol}}$ under $\rho_{\text{monop}} \leq 0.10$	Compliant bilateral agreement falls below 75% or any accepted patch violates $\mathcal{P}^* \sqsupseteq \mathcal{P}_U$
P4. Validator false-alarm containment	Schema canonicalization keeps boundary validator false alarms below 6% (vs. 37% in Maat [14])	Execute 1,000 benign compliant tool workflows under ActGov/ClawGuard boundary monitors [36, 37]	Benign compliant tool calls halted by validator false positives exceed 6.0%
P5. SD-JWT and ZK-PCC latency (Eq. 3)	Client-side lattice repair and SD-JWT/Bulletproofs attestation add < 50 ms per handshake	Benchmark Stage 2 + Stage 4 across 5,000 handshakes using RFC 9901 SD-JWT [16] and Bulletproofs [5]	95th-percentile negotiation and proof-generation latency exceeds 50.0 ms on laptop CPU
P6. Monopoly refusal break-even (ρ^*_{monop})	Statutory GPC/IEEE 7012 enforcement keeps live counterparty refusal below $\rho^*_{\text{monop}} = 0.455$	Crawl 1,000 commercial endpoints in CCPA/GDPR jurisdictions proffering `Sec-GPC: 1` and IEEE 7012 headers [38]	Unconditional service refusal or silent header stripping exceeds 45.5% in regulated jurisdictions

9 Conclusion

Autonomous personal agents cannot safeguard human privacy if they participate in the legacy notice-and-consent economy by parsing ambiguous legal prose inside an unconfined context window. Empirical evidence from ConFaide [24], AirGapAgent [4], and AgentLeak [13] demonstrates that unguarded and prompt-governed agents routinely violate Contextual Integrity norms and leak private attributes across internal channels. By synthesizing IEEE Std 7012-2025 first-party machine-readable privacy terms [19] with minimum-perturbation RFC-6902 lattice repair, explicit counterparty reservation-utility modeling [27], out-of-context boundary monitors [10, 37], and deterministic SD-JWT/ZK compliance proofs [16, 17], ZK-IEEE7012-PCC reduces simulated trace privacy exposure at 4-clause complexity from 93.33% [91.05%, 95.07%] (clickwrap) and 54.00%

[50.00%, 57.95%] (free-text LLM negotiation) to 0.00% [0.00%, 0.64%] under exact AST lowering (3.67% [2.43%, 5.49%] under $\epsilon_{\text{lower}} = 0.012$) while lifting compliant agreements from 5.50% [3.94%, 7.62%] (static matching) to 78.33% [74.86%, 81.44%]. At the same time, our lowering-fault (ϵ_{lower}) and monopoly refusal ($\rho^*_{\text{monop}} \approx 40.4\%$) analyses underscore that runtime compliance guarantees depend critically on policy-to-field compiler fidelity and statutory enforceability.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Faruk Alpay and Taylan Alpay. AgentSecBench: Measuring prompt injection, privacy leakage, and tool-use integrity in LLM agents. arXiv preprint arXiv:2605.26269, 2026. URL https://arxiv.org/abs/2605.26269.

[2] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local differential privacy on metric spaces: Optimizing the trade-off with utility. IEEE Computer Security Foundations Symposium (CSF), pages 262–276, 2018. URL https://doi.org/10.1109/CSF.2018.00026.

[3] Prerna Arote, Adrian Saiz, Oriol Saguillo, and Lucianna Kiffer. Cryptographic security is not enough: Privacy gaps in the Renegade decentralized dark pool. arXiv preprint arXiv:2609.27100, 2026. URL https://arxiv.org/abs/2609.27100.

[4] Eugene Bagdasarian, Ren Yi, Sahar Ghaleb, et al. AirGapAgent: Protecting privacy-conscious conversational agents. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024. URL https://arxiv.org/abs/2405.05175.

[5] Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. Bulletproofs: Short proofs for confidential transactions and more. IEEE Symposium on Security and Privacy (S&P), pages 315–334, 2018. URL https://doi.org/10.1109/SP.2018.00020.

[6] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[7] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[8] Lorrie Faith Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The platform for privacy preferences 1.0 (P3P1.0) specification. W3C Recommendation, 2002. URL https://www.w3.org/TR/P3P/.

[9] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2024. URL https://arxiv.org/abs/2406.13352.

[10] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[11] Dorothy E. Denning. A lattice model of secure information flow. Communications of the ACM, 19(5):236–243, 1976. URL https://doi.org/10.1145/360051.360056.

[12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. URL https://doi.org/10.1561/0400000042.

[13] Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. AgentLeak: A benchmark for internal-channel privacy leakage in multi-agent LLM systems. arXiv preprint arXiv:2602.11510, 2026. URL https://arxiv.org/abs/2602.11510.

[14] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[15] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Counting on dummies: Privacy amplification with observable counts and truncated batches. arXiv preprint arXiv:2610.42139, 2026. URL https://arxiv.org/abs/2610.42139.

[16] Daniel Fett, Kristina Yasuda, and Brian Campbell. Selective disclosure for JWTs (SD-JWT). IETF RFC 9901, 2025. URL https://datatracker.ietf.org/doc/rfc9901/.

[17] Matteo Frigo and abhi shelat. Anonymous credentials from ECDSA (Longfellow-ZK). IACR Communications in Cryptology, 1(2), 2024. URL https://eprint.iacr.org/2024/2010.

[18] Joseph A. Goguen and José Meseguer. Security policies and security models. IEEE Symposium on Security and Privacy, pages 11–20, 1982. URL https://doi.org/10.1109/SP.1982.10014.

[19] IEEE Standards Association / P7012 Working Group. Ieee std 7012-2025: Standard for machine readable personal privacy terms. IEEE Standards Association, 2025. URL https://standards.ieee.org/ieee/7012/7542/.

[20] IEEE. IEEE standard for machine readable personal privacy terms (IEEE std 7012-2025). Technical report, IEEE Standards Association, 2026. URL https://www.mystandards.biz/standard/ieee-7012-2025-20.1.2026.html. Published 20 January 2026.

[21] Juhee Kim and Woohyuk Choi. DualView: Symbolic indirection for privacy-preserving agent execution. arXiv preprint arXiv:2607.03821, 2026. URL https://arxiv.org/abs/2607.03821.

[22] Chen Liang and Fasheng Xu. When LLM agents negotiate: Private information and dynamic bargaining in supply chains. arXiv preprint arXiv:2608.07538, 2026. URL https://arxiv.org/abs/2608.07538.

[23] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark patterns at scale: Findings from a crawl of 11k shopping websites. Proceedings of the ACM on Human-Computer Interaction (CSCW), 3:81, 2019. URL https://doi.org/10.1145/3359183.

[24] Niloofar Mireshghallah, Hyunwoo Kim, Ximing Xenos, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.17884.

[25] Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. Journal of Economic Theory, 29(2):265–281, 1983. URL https://doi.org/10.1016/0022-0531(83)90048-0.

[26] Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–158, 2004. URL https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/.

[27] Sören Preibusch. Implementing privacy negotiations in e-commerce. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), LNCS 3918, pages 604–615, 2006. URL https://doi.org/10.1007/11731139_70.

[28] Tianneng Shi, Jingxuan He, Zhun Wang, Linyu Wu, Hongwei Li, Wenbo Guo, and Dawn Song. Progent: Programmable privilege control for LLM agents. arXiv preprint arXiv:2504.11703, 2025. URL https://arxiv.org/abs/2504.11703.

[29] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. URL https://doi.org/10.1142/S0218488502001648.

[30] Wei Wang, Abdul Hyee Waqas, and Burns Smith. Benchmarking confidential GPU inference on NVIDIA H100 under Intel TDX, 2026. URL https://arxiv.org/abs/2607.19353.

[31] Yutian Wang and Luyao Zhang. Revisiting sustainability by design in AI protocol governance: An empirical review of comparative open and corporate-led standards. arXiv preprint arXiv:2606.05667, 2026. URL https://arxiv.org/abs/2606.05667.

[32] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. The creation and analysis of a website privacy policy corpus (OPP-115). Proceedings of the Association for Computational Linguistics (ACL), pages 1330–1340, 2016. URL https://aclanthology.org/P16-1126/.

[33] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. IsolateGPT: An execution isolation architecture for LLM-based agentic systems. Network and Distributed System Security Symposium (NDSS), 2025. URL https://arxiv.org/abs/2403.04960.

[34] Wujie Xiong, Rabimba Karanjai, Yang Lu, Weidong Shi, and Lei Xu. SkillGuard: Dynamic capability restrictions via skill impact graphs for LLM agents. arXiv preprint arXiv:2608.30041, 2026. URL https://arxiv.org/abs/2608.30041.

[35] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. Findings of the Association for Computational Linguistics (ACL), 2024. URL https://arxiv.org/abs/2403.02691.

[36] Kaiyuan Zhang, Yuke Peng, Ke Jiang, and Yinqian Zhang. ActGov: Governing LLM agent actions via policy-constrained validation. arXiv preprint arXiv:2609.24446, 2026. URL https://arxiv.org/abs/2609.24446.

[37] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.

[38] Sebastian Zimmeck, Oliver Wang, Kuba Alicki, Jocelyn Wang, and Sophie Eng. Usability and enforceability of global privacy control. Proceedings on Privacy Enhancing Technologies (PoPETs), 2023(2):265–281, 2023. URL https://petsymposium.org/popets/2023/popets-2023-0051.php.