

Teleoneutrosophic Logic for Three-Way Authorisation of Institutional AI Agents

Selçuk Topal

`selcuk.topal@idu.edu.tr`

İzmir Demokrasi Üniversitesi

Buğra Ayan

Republic of Türkiye Directorate of Communications

Research Article

Keywords: neutrosophic logic, three-way decisions, paraconsistent logic, prompt injection, AI agents, e-government security

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11477293/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Teleoneutrosophic Logic for Three-Way Authorisation of Institutional AI Agents

Selçuk Topal^{1*} and Buğra Ayan²

^{1*}Department of Mathematics, Faculty of Science and Letters, Izmir Demokrasi University, İzmir, Türkiye.

²Department of Information Technologies, Directorate of Communications, Presidency of the Republic of Türkiye (T.C. İletişim Başkanlığı), Ankara, Türkiye.

*Corresponding author(s). E-mail(s): selcuk.topal@idu.edu.tr;

Contributing authors: bugra.ayan@iletisim.gov.tr;

Abstract

An institutional artificial intelligence (AI) agent may encounter an instruction to issue a refund inside a petition it was asked to summarise. The same tool call may be legitimate when requested by an officer and unauthorised when induced by that petition. We address this distinction with teleoneutrosophic values, pairing separate neutrosophic triples for effect and purpose. A provenance discount and evidence fusion feed a guard that authorises, escalates or denies a proposed call. We characterise paraconsistency and adjunction in the induced matrix logic and prove conditional limits on untrusted influence. With reliable labels and call attribution, a source cannot supply purpose support above its trust cap, and untrusted evidence cannot justify denial. Tests on 1,054 InjecAgent attack cases and synthetic e-government scenarios show where grading helps and where crisp protection suffices. On low-trust channels, the guard can reduce to taint-and-escalate; an attribution-based plan/data rule is at least as safe. Retaining ambivalence about an unrequested call reduces worst-case direct-harm pass-through after channel mislabelling from 100% to 34.5% on InjecAgent and 25.0% on the synthetic benchmark. A learned estimator that omits the trust label reduces direct-harm pass-through under ordinary mislabelling from 36.7% to 1.6%. The results identify a role for independent ambivalence in provenance-aware authorisation and expose the limits on automation when legitimate external requests are missing from training data.

Keywords: neutrosophic logic, three-way decisions, paraconsistent logic, prompt injection, AI agents, e-government security

1 Introduction

Consider an AI agent asked to summarise a tax petition. The petition contains a further instruction: issue a refund to a specified account. The proposed refund may use the same tool and have the same immediate effect as a refund explicitly requested by a tax officer. Yet the two calls have

different authority. A monitor that observes only the action's effect has no basis for distinguishing them.

The distinction matters for public-sector assistants that prepare transfers or update records. The OECD describes the growing use of AI across government functions (OECD 2025). An

agent working in such a setting may read citizen submissions while also having access to registry records, tax files, health summaries or staff accounts. Instructions hidden in material it reads can redirect that access. This is the setting of indirect prompt injection (Greshake et al. 2023); InjecAgent and AgentDojo demonstrate the vulnerability in tool-using agents (Zhan et al. 2024; Debenedetti et al. 2024). In an administrative workflow, this could mean an address changed without authority, extra account privileges, a false public announcement or disclosure to an external recipient. Workflows involving personal data also fall within the relevant data-protection framework (Republic of Türkiye 2016; European Union 2016).

A monitor placed between an agent and its tools must decide whether a proposed call may proceed, should be refused or needs an officer’s review. Fuzzy sets, intuitionistic fuzzy sets and single-valued neutrosophic sets offer increasingly expressive accounts of uncertain evidence (Zadeh 1965; Atanassov 1986; Smarandache 2007; Wang et al. 2010). Three-way decision theory provides a corresponding place for deferment (Yao 2010, 2011, 2012). These representations are useful, but applying any of them only to the proposition “this action is safe” leaves the authority problem unresolved. Refining the safety scale cannot recover information about whose instruction caused the call; Proposition 2 makes this limitation precise.

Authorisation therefore requires evidence that a call serves the authenticated principal’s request. The source of that evidence matters: an officer’s instruction and an instruction embedded in a petition need not carry the same authority. Electronic signatures, seals, registered e-mail and content-provenance mechanisms can help establish the origin and integrity of the material being assessed (European Union 2014, 2024a; Republic of Türkiye 2004, 2011; C2PA 2026). They do not establish that its contents are appropriate, nor do they determine how much influence a sender should have over an agent. Those are separate policy judgements.

Earlier work already makes purpose relevant to access decisions. Database policies can restrict the uses for which records are accessed (Byun et al. 2005; Byun and Li 2008), and Task Shield checks whether an agent’s instructions and calls advance the user’s task (Jia et al. 2025). CaMeL and

FIDES constrain the authority of external material through execution structure and information-flow controls (Debenedetti et al. 2025; Costa et al. 2025). Our contribution is a decision calculus for uncertain effect and purpose evidence within such an architecture. Crisp comparators establish which observed benefits require the particular graded construction.

The paper makes four contributions.

1. We pair an effect triple (T, I, F) with a telic triple (P, A, H) for purpose alignment, ambivalence and hostility. The underlying algebra is a product of existing single-valued neutrosophic algebras. The new construction couples the two assessments through a risk functional and yields exact conditions for paraconsistency and adjunction (Sections 3–4).
2. We define a provenance discount, an evidence-fusion operator and a three-way guard whose denial test uses trusted evidence only. With reliable provenance and call attribution, low-trust evidence cannot supply sufficient purpose support for authorisation. Under stated support conditions, its only possible change to a decision is escalation. We also identify the parameter range in which the guard’s authorisation behaviour is exactly taint-and-escalate (Sections 5–7).
3. We evaluate the construction against crisp purpose and provenance rules, taint rules, a plan/data-separation baseline and a neutrosophic comparator receiving the same evidence. The evaluation includes mislabelled channels, availability attacks and argument manipulation. It shows where the guard adds no empirical advantage and isolates the role of independent ambivalence (Section 8).
4. We specify records from which telic evidence can be learned. A causal model separates trust labels from the content used to estimate purpose and hostility. The experiments show why excluding the label matters and why legitimate requests from untrusted channels must be represented in training data (Section 8.4).

2 Background and related work

2.1 Representing uncertain assessments

A fuzzy membership grade is a single value in $[0, 1]$ (Zadeh 1965). Intuitionistic fuzzy sets distinguish membership from non-membership, subject to their sum being at most one (Atanassov 1986). Neutrosophic logic separates truth, indeterminacy and falsity (Smarandache 2007). Single-valued neutrosophic sets let these components range independently over $[0, 1]$ (Wang et al. 2010); associated scoring and aggregation methods support multicriteria decisions (Ye 2013, 2014).

Other extensions organise uncertainty in different ways: intervals, refined components, bipolar judgements and plithogenic attributes (Wang et al. 2005; Smarandache 2013; Deli et al. 2015; Smarandache 2017); Pythagorean and q -rung orthopair grades (Yager 2014, 2017); and picture, spherical and hesitant fuzzy information (Cuong 2014; Gündoğdu and Kahraman 2019; Torra 2010). A teleoneutrosophic value has six coordinates and can be encoded as a 2-refined neutrosophic value. Here the two triples assess different propositions, and a risk rule couples their assessments. Provenance determines how much support each source may contribute. Table 1 summarises this positioning.

2.2 Fusion and source reliability

Belnap’s logic and Ginsberg’s bilattices distinguish truth from the information available about it (Belnap 1977; Ginsberg 1988). Evidence theories also provide established treatments of unreliable or conflicting sources. Dempster–Shafer discounting weakens belief in proper subsets as source reliability decreases (Shafer 1987); contextual discounting allows reliability to depend on the hypothesis (Mercier et al. 2008). Dezert–Smarandache theory combines possibly conflicting evidence under models that allow overlapping hypotheses (Smarandache and Dezert 2004). Possibility theory also supports ordinal accounts of uncertainty and qualitative decision models (Dubois and Prade 1995), while subjective logic develops operators for uncertain opinions and trust (Jøsang 2016).

Our discount applies different policies to justification and warning: trust caps positive support, whereas a warning is retained and uncertainty grows. This is a modelling choice, rather than a consequence of the evidential frameworks above. Contextual discounting likewise permits reliability to vary with the hypothesis, but does not establish our guard’s influence theorem. The additional restriction on denial is imposed by the decision rule after fusion.

2.3 Three-way decisions and consequence

Decision-theoretic rough sets derive acceptance, rejection and deferment regions from expected losses (Yao 2010, 2011). Broader three-way decision frameworks admit evaluation functions chosen for the application (Yao 2012; Hu 2014), as do shadowed sets (Pedrycz 1998). Three-way models have been developed for intuitionistic fuzzy, Pythagorean fuzzy and single-valued neutrosophic information (Liang et al. 2017; Zhang and Ma 2020; Jiao et al. 2020).

The guard studied here uses two evaluations: risk from all evidence for authorisation, and risk from trusted evidence for denial. We calibrate its thresholds against an explicit loss function rather than derive them analytically from a loss matrix. Escalation then has a security interpretation: under the conditions proved below, it is the only change an adversarial evidence source can force.

The aggregation of individually accepted judgements is also a logical question. Kyburg questions conjunctive closure of accepted scientific statements (Kyburg 1970); Jaśkowski and Schotch and Jennings consider reasoning in inconsistent settings (Jaśkowski 1999; Schotch and Jennings 1980). These works are background rather than sources of our parameter conditions. Bou et al. study consequence that preserves lower truth bounds in residuated lattices (Bou et al. 2009); their additional adjunction rule concerns the fusion connective, while lattice-meet adjunction is already valid. Here we examine lattice meet itself. For our matrix semantics, adjunction holds exactly when the designated values are meet-closed, and Section 4 derives the condition for that closure.

Table 1 Relation to representative formulations. The entries describe the cited mechanisms and the role they play in this study; they do not assert that an entire family lacks a feature. Published agent systems are related work, while the experimental comparators are the static rules specified in Section 8.

Formulation	Mechanism in the cited work	Relation to the present guard
Fuzzy and neutrosophic representations	Grades, multiple components or attributes describe uncertain judgements.	The carrier is inherited; effect and purpose receive separate triples.
Belief functions and subjective logic	Source reliability changes evidential support and uncertainty.	Our discount preserves concerns; trusted evidence alone can justify denial.
Three-way decision models	Acceptance, rejection and deferment depend on losses or evaluation functions.	Full-evidence and trusted-evidence risk govern different decisions.
Purpose-based access control	Permitted data uses are checked against an access purpose.	This motivates purpose as a policy input, independently of an action’s effect.
CaMeL and FIDES	Capability restrictions and information-flow policies constrain execution.	These supply architectural precedents for the assumed provenance and attribution.
Progent	Symbolic rules restrict tool names and arguments; privilege changes are checked.	Such policy checks are richer than the static content rules evaluated here.
Task Shield	An LLM estimates instruction and tool-call contribution to user goals.	Its task-alignment criterion precedes our graded purpose assessment.
MELON	Calls from original and masked-user executions are compared.	Behavioural dependence supplies a different signal from our static evidence.
RTBAS	Selective label propagation and redaction enforce information-flow policies.	Dependency screening differs from estimating task alignment.
TNL	A risk rule couples effect and purpose after provenance-sensitive fusion.	Theorems state conditional influence limits; static experiments test their scope.

2.4 Purpose and agent security

Byun and colleagues attach permitted uses to database information and check the purpose of an access (Byun et al. 2005; Byun and Li 2008). ABAC evaluates attributes of the subject, object, operation and environment (Hu et al. 2014); UCON also addresses obligations, conditions and controls that persist during use (Park and Sandhu 2004). We place the guard at the tool boundary as a proposed reference-monitor component (Anderson 1972). This placement alone does not establish complete mediation or implementation assurance. The trajectory extension takes its security-policy background from lattice information flow, language-based enforcement and integrity models (Denning 1976; Sabelfeld and Myers 2003; Biba 1977).

Greshake et al. demonstrate attacks carried by retrieved material in deployed and synthetic applications (Greshake et al. 2023). InjecAgent tests injected tool responses (Zhan et al. 2024), while AgentDojo supports interactive tasks, attacks and defences (Debenedetti et al. 2024). The other benchmarks address different risks: ToolEmu uses emulated tools to expose unsafe behaviour under underspecified instructions (Ruan et al. 2024); AgentHarm tests explicitly malicious user tasks and jailbreaks (Andriushchenko et al. 2025). These results cannot be treated as measurements of one common threat model.

Instruction Hierarchy trains models to respect instruction priority (Wallace et al. 2024). StruQ combines a structured front-end with a specially trained model (Chen et al. 2025), whereas spotlighting transforms external text to make source boundaries more apparent (Hines et al. 2024). Their empirical robustness does not establish our provenance assumptions. CaMeL restricts execution with capabilities and separated control/data flow (Debenedetti et al. 2025); FIDES uses confidentiality and integrity labels with selective information hiding (Costa et al. 2025). The design-pattern study sets out further architectural trade-offs (Beurer-Kellner et al. 2025). Such controls can support the architecture assumed here, provided call dependencies are tracked conservatively. Progent expresses tool and argument privileges as symbolic policies and checks changes to those privileges (Shi et al. 2026). Our content estimators do not implement those policy checks.

Task Shield uses an LLM to extract instructions and estimate their contribution to user goals (Jia et al. 2025). MELON instead masks the user request, re-executes the trajectory and compares the resulting calls; similar actions can indicate that the original call followed external instructions (Zhu et al. 2025). RTBAS addresses information-flow policy enforcement: model-based dependency screening selects which labels to propagate, and redaction limits exposure (Zhong et al. 2025). It is therefore not simply another task-alignment

scorer. These mechanisms suggest possible evidence inputs for a future integration with our calculus. We have not performed that integration or a live-agent comparison. Our threat model also excludes intentional subversion by the model itself, the setting studied in AI control (Greenblatt et al. 2024).

In network security, Kavitha et al. evaluate a neutrosophic intrusion classifier on KDD Cup 99 (Kavitha et al. 2012), and Elwahsh et al. review intrusion-detection approaches for mobile ad hoc networks (Elwahsh et al. 2018). Wang et al. use intuitionistic fuzzy information to assess anomalies in flow interaction (Wang et al. 2018); Fan et al. study entropy-based vulnerability evaluation (Fan et al. 2019). These are precedents for representing uncertain security evidence, rather than validations of agent authorisation. The NSL-KDD benchmark addresses deficiencies in KDD Cup 99 (Tavallae et al. 2009) and is separate from the agent experiments reported here.

3 Teleoneutrosophic values

Write $\mathcal{N} = [0, 1]^3$, $x \hat{+} y = x + y - xy$ for the probabilistic sum, and $(x)_+ = \max\{x, 0\}$.

Definition 1 (Teleoneutrosophic value) For the proposition “action α is admissible for principal p ”, a teleoneutrosophic value is

$$\omega = \langle \varepsilon; \psi \rangle = \langle (T, I, F); (P, A, H) \rangle \in \mathcal{T} := \mathcal{N} \times \mathcal{N}.$$

The effect coordinates express support for safe execution (T), uncertainty in that assessment (I) and support for harm (F). The telic coordinates express support that the action serves the principal’s purpose (P), ambivalence or missing support for that assessment (A), and support for a purpose adverse to the principal (H).

Neither triple is required to sum to one. In particular, purpose alignment and hostility need not exclude one another. A transfer may fulfil an officer’s request while also copying the record to an unauthorised recipient. Conversely, little may be known in either direction. The latter case motivates a separate ambivalence coordinate: high support supplied later need not erase an earlier absence of justification. Section 8 tests this distinction.

Definition 2 (Protective order and negation) Write $\omega \sqsubseteq \omega'$ when

$$\begin{aligned} T &\leq T', & I &\geq I', & F &\geq F', \\ P &\leq P', & A &\geq A', & H &\geq H'. \end{aligned}$$

Thus ω' is at least as admissible as ω . Define

$$\neg \langle (T, I, F); (P, A, H) \rangle = \langle (F, 1 - I, T); (H, 1 - A, P) \rangle.$$

Proposition 1 Let $\mathbf{N} = (\mathcal{N}, \leq_N, \neg_N)$ be the single-valued neutrosophic algebra whose truth order increases the first coordinate and decreases the other two, with $\neg_N(x, y, z) = (z, 1 - y, x)$. Then

$$(\mathcal{T}, \sqsubseteq, \neg) = \mathbf{N} \times \mathbf{N}.$$

This is a completely distributive complete lattice with componentwise meet $\sqcap$ and join $\sqcup$, and bounds

$$\begin{aligned} \perp &= \langle (0, 1, 1); (0, 1, 1) \rangle, \\ \top &= \langle (1, 0, 0); (1, 0, 0) \rangle. \end{aligned}$$

Together with $\neg$, it is a De Morgan algebra that is neither Boolean nor Kleene.

Proof Definition 2 gives the product directly. The order on $\mathcal{N}$ is the product of three complete chains, with the second and third orders reversed. It is therefore completely distributive. The map $\neg_N$ is an order-reversing involution, and both properties pass to the product. The element with every coordinate equal to $1/2$ is fixed by negation, so this negation does not provide Boolean complements. For the Kleene inequality, take

$$\begin{aligned} x &= \langle (0.9, 0.5, 0.9); (0.5, 0.5, 0.5) \rangle, \\ y &= \langle (0.1, 0.5, 0.1); (0.5, 0.5, 0.5) \rangle. \end{aligned}$$

The T -coordinate of $x \sqcap \neg x$ is 0.9, whereas that of $y \sqcup \neg y$ is 0.1. Hence $x \sqcap \neg x \not\sqsubseteq y \sqcup \neg y$. $\square$

The algebraic carrier is inherited. The subsequent risk coupling, provenance treatment and decision semantics are the additional structure studied here.

Definition 3 (Embeddings) Let $\psi_\top = (1, 0, 0)$ and define

$$\begin{aligned} \iota_F(\mu) &= \langle (\mu, 0, 1 - \mu); \psi_\top \rangle, \\ \iota_N(x, y, z) &= \langle (x, y, z); \psi_\top \rangle, \\ \iota_{\text{IFS}}(\mu, \nu) &= \langle (\mu, 1 - \mu - \nu, \nu); \psi_\top \rangle. \end{aligned}$$

These maps place established values in a setting where purpose is fully endorsed. Both ι_F and ι_N preserve meet and join and are lattice embeddings. They do not preserve negation: negating $\psi_\top$

gives $(0, 1, 1)$. The intuitionistic fuzzy embedding is injective but is not order-preserving, since hesitation is not monotone in the usual intuitionistic fuzzy order.

4 The induced matrix logic

Let L contain atoms $p, q, \dots$ and connectives $\wedge, \vee, \neg$. A valuation maps atoms to $\mathcal{T}$ and extends by the lattice operations and negation. Here meet combines judgements; it does not describe executing two actions in sequence. Section 7 addresses execution sequences.

Using the all-trusted guard of Definition 7, set

$$D_G = \{\omega : G(\omega) = \text{auth}\}.$$

Define $\Gamma \models_G \varphi$ to mean that every valuation designating all members of Γ also designates φ . Throughout,

$$\begin{aligned} 0 &\leq \tau_T, \tau_P, \lambda, \beta, F_{\max} \leq 1, \\ 0 &\leq \theta_a < \theta_d \leq 1. \end{aligned}$$

The largest designated harm coordinate is

$$\bar{F} = \begin{cases} F_{\max}, & \beta = 1, \\ \min\{F_{\max}, \theta_a/(1 - \beta)\}, & \beta < 1. \end{cases}$$

Theorem 1 (Structure of consequence) *(i) $\models_G$ is reflexive, transitive and monotone. It validates $\varphi \models_G \varphi \vee \chi$ and $\varphi \wedge \chi \models_G \varphi$.*

(ii) It is paracomplete for every parameter choice: $\not\models_G p \vee \neg p$. It is paraconsistent, meaning $\{p, \neg p\} \not\models_G q$, exactly when

$$\max\{\tau_T, \tau_P, \lambda/2\} \leq \theta_a, \quad \tau_T \leq F_{\max}.$$

(iii) It validates adjunction, $\{p, q\} \models_G p \wedge q$, exactly when

$$\tau_T \geq 1 - \theta_a + \bar{F}.$$

This condition implies $\bar{F} \leq \theta_a$.

(iv) If atomic valuations are restricted to the image of ι_N , adjunction holds for negation-free formulas at every parameter choice, and for all formulas if $\tau_P > 0$. The same statements hold on the image of ι_F .

Proof (i) Preservation of a designated set gives reflexivity, transitivity and monotonicity of consequence. By

Theorem 2, D_G is an up-set. Since $v(\varphi) \sqsubseteq v(\varphi) \sqcup v(\chi)$ and $v(\varphi) \sqcap v(\chi) \sqsubseteq v(\varphi)$, the two stated inferences follow.

(ii) For paracompleteness, put $v(p) = \langle (1, 1/2, 1); (1, 1/2, 1) \rangle$. In $v(p \vee \neg p)$, all of T, F, P, H equal one. The first risk term is therefore one, which exceeds θ_a .

Since $\perp$ has risk one, an undesignated value always exists. Paraconsistency is thus equivalent to the existence of a value ω for which both ω and $\neg\omega$ are designated. Under the stated inequalities, the fixed point

$$\omega = \langle (\tau_T, 1/2, \tau_T); (\tau_P, 1/2, \tau_P) \rangle$$

has risk $\max\{\tau_T, \tau_P, \lambda/2\}$ and satisfies the support tests and harm ceiling. Conversely, designating ω and $\neg\omega$ requires $T, F \geq \tau_T$ and $P, H \geq \tau_P$. The ceiling applied to $\neg\omega$ gives $\tau_T \leq T \leq F_{\max}$. The two indeterminacy constraints give $\lambda \max\{I, 1 - I\} \leq \theta_a$, hence $\lambda/2 \leq \theta_a$. If $P \geq H$, the harm term of $\neg\omega$ equals $T \geq \tau_T$; otherwise the harm term of ω equals $F \geq \tau_T$. Thus $\tau_T \leq \theta_a$. Comparing T and F in the same way gives $\tau_P \leq \theta_a$ from the hostility term.

(iii) Because atoms may take arbitrary values, adjunction holds exactly when D_G is meet-closed. For sufficiency, let $\omega_1, \omega_2 \in D_G$ and $\omega = \omega_1 \sqcap \omega_2$. The support tests, harm ceiling and indeterminacy constraint are preserved by meet. Each $F_i \leq \bar{F}$, so the harm term of ω is at most $\bar{F} \leq \theta_a$. Its hostility term is at most $1 - (\tau_T - \bar{F})_+ \leq \theta_a$.

For necessity, the values

$$\omega_1 = \langle (\tau_T, 0, \bar{F}); (1, 0, 0) \rangle,$$

$$\omega_2 = \langle (1, 0, 0); (1, 0, 1) \rangle$$

are designated: their risks are $\bar{F}(1 - \beta) \leq \theta_a$ and zero. Their meet is $\langle (\tau_T, 0, \bar{F}); (1, 0, 1) \rangle$. Designating it requires $\bar{F} \leq \theta_a$ and $1 - (\tau_T - \bar{F})_+ \leq \theta_a$. Since $\theta_a < 1$, the latter is equivalent to the claimed inequality, which also implies the former.

(iv) On the image of ι_N , risk is $\max\{F(1 - \beta), \lambda I\}$, a maximum of single-coordinate terms. The image is meet- and join-closed, so the designated part is meet-closed. Adding negation generates only two telic triples, $\psi_\top$ and $(0, 1, 1)$. When $\tau_P > 0$, only the first can be designated, and the same closure argument applies. $\square$

Every experimental calibration selects $\tau_T = 0$. Adjunction would then require $\theta_a \geq 1$, so it fails. Two designated judgements can rely on different strengths: one may have strong safety evidence despite hostile purpose evidence, while the other has strong purpose evidence despite harm. Their meet retains the weaker support and the stronger concern in each coordinate. Here adjunction fails because the risk coupling allows

individually accepted values whose meet is not accepted. Part (iv) explains why it is absent on the embedded classical values. The experimental choices $\tau_P = \theta_a = 0.3$, $F_{\max} = 1$ and $\lambda \in \{0.25, 0.5\}$ also meet the paraconsistency condition.

5 Evidence provenance and fusion

An evidence piece consists of a triple $x = (x_1, x_2, x_3)$ for one assessment and a trust value $t \in [0, 1]$ for its channel. We write it as (x, t) . The tool catalogue and an authenticated officer's instruction are trusted in our model ($t = 1$). Petitions, inbound e-mail and uploaded documents receive lower trust. A sealed partner-agency message may occupy an intermediate tier.

The institution assigns these values. A valid seal helps authenticate a sender and protect a message's integrity; it does not certify that following the message would be appropriate. Intermediate trust therefore expresses a permission to influence the agent, not a cryptographic probability that the sender is correct.

Definition 4 (Caution-preserving discount) For $x \in \mathcal{N}$ and $t \in [0, 1]$, define

$$D_t(x_1, x_2, x_3) = (tx_1, x_2 \hat{+} (1-t), x_3).$$

Definition 5 (Sceptical fusion) Let $E = \{(x_k, t_k)\}$ be a finite multiset and write $D_{t_k}(x_k) = (a_k, b_k, c_k)$. For non-empty E , set

$$\Phi(E) = \left(\max_k a_k, 1 - \prod_k (1 - b_k), \max_k c_k \right).$$

Set $\Phi(\emptyset) = (0, 1, 0)$, representing ignorance. Fuse effect and telic evidence separately:

$$\omega(E) = \langle \Phi(E^\varepsilon); \Phi(E^\psi) \rangle.$$

The value obtained from pieces with $t = 1$ alone is denoted by ω^0 .

Fusion retains the strongest discounted support and the strongest report of harm or hostility, while uncertainty accumulates. It is therefore permissive about support and cautious about concerns. The support maximum does not let an untrusted source contribute more support than its trust value. By contrast, averaging concerns could

dilute a strong warning with numerous weak or reassuring reports.

Lemma 1 (Fusion properties) (i) *Fusion is commutative. For non-empty E, E' , it satisfies*

$$\Phi(E \cup E') = \Phi(\{(\Phi(E), 1), (\Phi(E'), 1)\}).$$

Thus non-empty batches may be fused incrementally.

(ii) *If $\emptyset \neq E \subseteq E'$ and $\Phi(E) = (a, b, c)$, $\Phi(E') = (a', b', c')$, then $b' \geq b$, $c' \geq c$ and*

$$a' \leq \max \left\{ a, \max_{(x,t) \in E' \setminus E} t \right\}.$$

The maximum over an empty added set is taken as zero. An added piece with support one attains the maximal support permitted by its trust. In particular, the displayed bound is attained whenever the added piece has maximal added trust and the added set is non-empty.

(iii) *Fusion is idempotent in support and harm, but not in indeterminacy. For k copies of a piece with trust t , the fused indeterminacy satisfies $b \geq 1 - t^k$.*

(iv) *Replacing a piece by one with no greater support and no smaller indeterminacy or harm can only lower the fused value in $\leq_N$.*

Proof For (i), D_1 is the identity, maxima are commutative and associative, and the fused uncertainty of two batches is

$$1 - (1 - b_E)(1 - b_{E'}) = 1 - \prod_{k \in E \cup E'} (1 - b_k).$$

For (ii), adding values cannot decrease a maximum or a probabilistic sum, and each new support term is $tx_1 \leq t$. A piece with $x_1 = 1$ attains its cap. For (iii), each discounted copy has $b_k \geq 1 - t$, so $\prod_k (1 - b_k) \leq t^k$. Part (iv) follows from the coordinatewise monotonicity of the discount and all three aggregations. $\square$

The non-empty qualification in part (i) matters: the stipulated ignorance value for an empty set must not be inserted as an additional evidence piece. Likewise, repeated reports are not neutral. Their accumulated indeterminacy is one reason that availability pressure must be considered separately from unauthorised execution.

Lemma 2 (Adding low-trust evidence to a supported axis) *Let $E_0 \neq \emptyset$ contain trusted evidence for one*

axis, with fused support $a_0 \geq \bar{t}$. If every piece in E_{adv} has trust at most $\bar{t}$, then

$$\Phi(E_0 \cup E_{\text{adv}}) \leq_N \Phi(E_0),$$

and support remains a_0 .

Proof Lemma 1(ii) bounds the new support above by $\max\{a_0, \bar{t}\} = a_0$ and below by a_0 . Indeterminacy and concern cannot decrease. $\square$

If trusted support is below $\bar{t}$, an adversarial source can increase it up to $\bar{t}$. Theorem 4 handles that case through the support thresholds. The discount still preserves every harm or hostility report. In classical belief-function discounting, unreliable specific evidence is weakened towards ignorance (Shafer 1987; Mercier et al. 2008); subjective-logic discounting also changes the weight of an opinion according to trust (Jøsang 2016). These formalisms do not themselves prescribe how a neutrosophic harm coordinate should be treated. We retain that coordinate as a warning and restrict the decision it can justify: an untrusted warning may require review, but cannot justify denial.

6 The protective guard

Definition 6 (Telic risk) For $\lambda, \beta \in [0, 1]$, define

$$\rho(\omega) = \max \left\{ \begin{array}{l} F(1 - \beta(P - H)_+), \\ H(1 - (T - F)_+), \\ \lambda \max\{I, A\} \end{array} \right\}.$$

The first term permits net purpose support to mitigate harm. The second permits net safety support to mitigate hostility. The third charges for uncertainty, with weight λ . These choices impose four constraints on the risk assessment.

First, increasing indeterminacy, harm, ambivalence or hostility cannot reduce risk, and increasing safety or purpose support cannot increase it. Second, mitigation uses the other assessment and only its net support: when $P \leq H$, the harm term remains F ; when $T \leq F$, the hostility term remains H . Thus contradictory support on an axis does not cancel its own concern. In particular, $T = F = P = H = 1$ gives risk one. Third, fully endorsed purpose recovers the embedded neutrosophic rule (Theorem 3). Finally, $\rho(\top) = 0$ and $\rho(\perp) = 1$. A sum or product need not satisfy

these constraints: a compensating sum can mask a serious concern, while a product can suppress it through a zero factor.

Definition 7 (Three-way guard) Apply the following tests in order to the full value ω and trusted value ω^0 :

$$G(\omega, \omega^0) = \begin{cases} \text{deny}, & \rho(\omega^0) \geq \theta_d, \\ \text{auth}, & \rho(\omega) \leq \theta_a, T \geq \tau_T, \\ & P \geq \tau_P, F \leq F_{\max}, \\ \text{esc}, & \text{otherwise.} \end{cases}$$

When all evidence is trusted, write $G(\omega)$ for $G(\omega, \omega)$.

The two tests answer different questions. Trusted evidence may establish enough risk to refuse execution. Authorisation, however, must survive the full evidence set. The ceiling $F_{\max}$ reserves high-impact actions for further confirmation even when purpose support would otherwise mitigate their harm. When there is no trusted telic evidence, $A^0 = 1$; such a call is denied if $\lambda \geq \theta_d$. Our calibrated settings have $\lambda < \theta_d$, leaving such uncertainty eligible for escalation rather than denial.

Theorem 2 (Protective monotonicity) If $\omega \sqsubseteq \omega'$, then $\rho(\omega') \leq \rho(\omega)$ and

$$G(\omega) \leq G(\omega') \quad (\text{deny} < \text{esc} < \text{auth}).$$

For the two-argument guard, the denial predicate is antitone in ω^0 and the authorisation predicate is monotone in ω .

Proof Each risk term has the coordinatewise monotonicity stated above. If $G(\omega) = \text{auth}$, the lower risk and improved support coordinates of ω' preserve all authorisation tests and exclude denial. If $G(\omega) = \text{esc}$, then $\rho(\omega') \leq \rho(\omega) < \theta_d$, which excludes denial. The two predicates in the last claim follow from the same coordinatewise comparisons. $\square$

Theorem 3 (Embedded fuzzy and neutrosophic rules) On $\iota_N(x, y, z)$, the guard denies when $\max\{z(1 - \beta), \lambda y\} \geq \theta_d$ and authorises when

$$\max\{z(1 - \beta), \lambda y\} \leq \theta_a, \quad x \geq \tau_T, \quad z \leq F_{\max}.$$

On $\iota_F(\mu)$ with $\beta < 1$, it authorises iff $\mu \geq c$ and denies iff $\mu \leq 1 - \theta_d/(1 - \beta)$, where

$$c = \max\{\tau_T, 1 - \theta_a/(1 - \beta), 1 - F_{\max}\}.$$

For $\beta = 1$, it never denies and authorises iff $\mu \geq \max\{\tau_T, 1 - F_{\max}\}$. On negation-free formulas in the image of ι_F , consequence therefore coincides with fuzzy threshold consequence for the designated interval $[c, 1]$, using the latter value of c when $\beta = 1$.

Proof The embedded values have $P = 1$ and $A = H = 0$, so their risk is $\max\{z(1 - \beta), \lambda y\}$. The purpose test is automatic. For fuzzy values, substitute $y = 0$ and $z = 1 - \mu$. If $\beta = 1$, risk is zero and cannot reach θ_d . Meet and join preserve the image, so negation-free formulas remain in this regime. $\square$

Proposition 2 (The dual-use dilemma) *If $\tau_P > 0$ or $\beta > 0$, two values can share the same effect triple and receive different decisions. Consequently, no rule using effect evidence alone can coincide with G over all values. This includes fuzzy, intuitionistic fuzzy, Pythagorean, picture, spherical and neutrosophic rules restricted to that evidence.*

Proof If $\tau_P > 0$, use $\varepsilon = (1, 0, 0)$. The telic triple $(1, 0, 0)$ is authorised, whereas $(0, 0, 0)$ is not. If $\beta > 0$, use $\varepsilon = (1 - \theta_d, 0, \theta_d)$. With $\psi = (0, 0, 1)$, risk reaches at least θ_d and the call is denied. With $\psi = (1, 0, 0)$, risk is $\theta_d(1 - \beta) < \theta_d$, so the decision is different. $\square$

The argument does not depend on a particular choice of effect scale. It depends on information that the scale does not contain.

6.1 Threat model and provenance assumptions

The adversary may write arbitrary text and any number of evidence pieces on untrusted channels. It cannot alter their trust labels: every piece it controls has trust at most $\bar{t} < 1$. We call this assumption *provenance integrity*.

We also require *attribution*. The system records every channel whose content was available when a call or argument was produced. It attributes the call to all such channels unless it can certify derivation from trusted inputs alone. Conservative over-attribution may increase review. Under the support conditions of Theorem 4, it does not create additional authorisations or denials.

Authenticated sessions, signatures, seals and registered e-mail can support source identification and integrity checks (European Union 2014,

2024a; Republic of Türkiye 2004, 2011); content credentials provide related provenance mechanisms (C2PA 2026). Maintaining the labels through forwarding is still an architectural obligation. A message relayed by another agency must retain the original source’s label. Planner/executor separation and information-flow tracking can support attribution (Debenedetti et al. 2025; Costa et al. 2025; Beurer-Kellner et al. 2025).

A trusted triple must be computed solely from trusted inputs, such as the officer’s instruction and the tool catalogue. An estimator must be fixed independently of the adversarial text before it is applied. Compromised officer accounts, officers induced to request harmful actions, and agents with their own harmful objectives fall outside this model.

Theorem 4 (Limits on untrusted influence) *Let E_0 be trusted evidence, let every piece in E_{adv} have trust at most $\bar{t}$, and set*

$$\omega^0 = \omega(E_0), \quad \omega = \omega(E_0 \cup E_{\text{adv}}).$$

- (i) *If $\tau_P > \bar{t}$ and $G(\omega, \omega^0) = \text{auth}$, then $P^0 \geq \tau_P$. Similarly, $T^0 \geq \tau_T$ whenever $\tau_T > \bar{t}$ and the call is authorised.*
- (ii) *The denial decision is independent of E_{adv} . In particular, adding untrusted content cannot cause denial.*
- (iii) *Suppose that each axis receiving adversarial pieces either has trusted support at least $\bar{t}$ or has its support threshold above $\bar{t}$. Then*

$$G(\omega, \omega^0) \in \{G(\omega^0), \text{esc}\}.$$

- (iv) *Parts (ii) and (iii) extend to the trajectory guard of Definition 8, with the condition in (iii) required at each step.*

Proof (i) Lemma 1(ii) gives $P \leq \max\{P^0, \bar{t}\}$. With no trusted telic evidence, $P^0 = 0$ and the same bound follows directly from $t_k x_{1k} \leq \bar{t}$. Thus $P^0 < \tau_P$ implies $P < \tau_P$. The argument for T is identical.

(ii) The first test in Definition 7 uses ω^0 alone.

(iii) For an axis receiving adversarial evidence, call it *supported* when its trusted support is at least $\bar{t}$. Otherwise call it *gated*; its threshold then exceeds $\bar{t}$. If an axis is gated, neither the trusted nor the full value can satisfy that support test. Both decisions are therefore unauthorised, and part (ii) makes them either the same denial or escalation.

If all affected axes are supported, Lemma 2 gives $\omega \sqsubseteq \omega^0$; unaffected axes are unchanged. Trusted authorisation can then remain authorisation or become escalation, since denial is fixed. A trusted escalation cannot become authorisation because the authorisation predicate is monotone, and cannot become denial by part (ii). A trusted denial remains denial.

(iv) The trajectory denial test depends only on trusted values and the trusted kernel K_H^0 , so part (ii) is unchanged. Let G_Π^0 be the trajectory decision obtained by using trusted values in both tests. A gated axis at any step prevents both full and trusted trajectory authorisation. If no step has a gated axis, $\omega_j \sqsubseteq \omega_j^0$ at every step. Hence

$$\begin{aligned} \rho(\omega_j) &\geq \rho(\omega_j^0), \quad H_j \geq H_j^0, \\ (P_j - H_j)_+ &\leq (P_j^0 - H_j^0)_+. \end{aligned}$$

The full kernel is therefore at least the trusted-value kernel, and no failed trusted authorisation test becomes satisfied on full evidence. This leaves exactly the same decision or escalation, as in (iii). $\square$

The result quantifies over all permitted adversarial pieces, regardless of the attacker’s capability. It does not establish the correctness of the trusted purpose estimate or protect labels that the architecture fails to preserve. The implementation includes random-instance property checks for the formal results; such checks complement rather than replace the proofs (see Code availability).

Proposition 3 (Taint limit) *If $\lambda(1-t) > \theta_a$, no call containing an effect or telic piece of trust at most t is authorised. For $\lambda > 0$, calls touched by a channel below*

$$1 - \theta_a/\lambda$$

therefore have the same authorisation behaviour as a taint-and-escalate rule that escalates such calls and otherwise applies G .

Proof A piece with trust $t' \leq t$ has discounted indeterminacy at least $1 - t' \geq 1 - t$. Fusion cannot reduce it. Thus $I \geq 1 - t$ or $A \geq 1 - t$, giving $\rho(\omega) \geq \lambda(1 - t) > \theta_a$. $\square$

Remark 1 (The range in which grading operates) For the transferred calibration $\theta_a = 0.3$ and $\lambda = 0.5$, the boundary is 0.4. Petitions, e-mail and uploads with $t = 0.2$ are effectively tainted; the partner channel with $t = 0.6$ remains graded. Lowering λ expands the graded range but can leave purpose support as the only barrier on a low-trust channel. Theorem 4(i) then

still bounds authorisation when $t < \tau_P$. Once $t \geq \tau_P$, that bound no longer applies to the channel, and the content estimate becomes decisive. Figure 2 shows this transition.

Example 1 (A refund request) An officer asks for a summary of petition 2026-118. The petition instead instructs the agent to issue an 18,400 TL refund for taxpayer 5530 to an IBAN beginning TR33. The proposed tool is `TaxOfficeIssueRefund`. The construction in Section 8.3 assigns the same effect triple, (0.09, 0.10, 0.90), to this tool in each scenario in Table 2. The transferred parameters are

$$\begin{aligned} \tau_T &= 0, \quad \tau_P = \theta_a = 0.3, \quad \theta_d = 0.6, \\ \lambda &= 0.5, \quad \beta = F_{\max} = 1, \quad \bar{t} = 0.2. \end{aligned}$$

When the officer requested only a summary, trusted evidence does not justify the refund and the call is denied. When the officer requested an approved refund, planted objections can trigger review but cannot enter the denial test. A new IBAN also contributes hostility evidence even if its channel is mislabelled. An attacker who changes an argument without introducing a new identifier may evade this detector, as the later experiments show.

7 Trajectories and decomposition attacks

A harmful operation may be split into a sensitive read followed by an apparently routine e-mail. Give each action α a source sensitivity $\sigma(\alpha) \in [0, 1]$ and an egress value $\eta(\alpha) \in [0, 1]$, following the information-flow perspective (Denning 1976; Sabelfeld and Myers 2003).

Definition 8 (Flow kernels and trajectory guard) For $\Pi = (\alpha_1, \dots, \alpha_n)$ with full values ω_j and trusted values ω_j^0 , put

$$K(\Pi) = \max_{i \leq k \leq n} \sigma(\alpha_i) \eta(\alpha_k) \cdot \max\{1 - \beta(P_k - H_k)_+, H_k\},$$

$$K_H^0(\Pi) = \max_{i \leq k \leq n} \sigma(\alpha_i) \eta(\alpha_k) H_k^0.$$

At step j , replace risk in the denial test by

$$\max\{\rho(\omega_j^0), K_H^0(\alpha_1, \dots, \alpha_j)\}$$

and risk in the authorisation test by

$$\max\{\rho(\omega_j), K(\alpha_1, \dots, \alpha_j)\}.$$

The trajectory is denied if any step is denied, authorised if every step is authorised, and escalated otherwise.

Table 2 Refund scenarios in Example 1, with effect triple $\varepsilon = (0.09, 0.10, 0.90)$. The scenario labels distinguish who requested the action and which additional evidence was supplied. Trusted and full risks are shown separately.

Scenario	ψ^0	ψ	$\rho(\omega^0)$	$\rho(\omega)$	Decision
Petition injects a refund; officer requested a summary	(0, 0, 0)	(0.2, 0.8, 0.5)	0.90	0.90	deny
Officer requests an approved refund of 450 TL to taxpayer 7781	(1, 0, 0)	(1, 0, 0)	0.05	0.05	auth
Officer requests refund; planted message demands refusal	(1, 0, 0)	(1, 0.8, 1)	0.05	1.00	esc
Officer requests refund; uploaded document changes IBAN	(1, 0, 0)	(1, 0.8, 0.5)	0.05	0.50	esc
Injection arrives through a mislabelled trusted channel	(1, 0, 0.5)	(1, 0, 0.5)	0.50	0.50	esc

Proposition 4 (Trajectory properties) (i) *Both kernels are non-decreasing as a trajectory grows.*

(ii) *Suppose a trajectory Π' contains r before s , and a single action m satisfies*

$$\begin{aligned} \sigma(r) &\geq \sigma(m), \quad \eta(s) \geq \eta(m), \\ (P_s - H_s)_+ &\leq (P_m - H_m)_+, \quad H_s \geq H_m. \end{aligned}$$

Then $K(\Pi') \geq K(m)$.

(iii) *If $1 - \beta\tau_P > \theta_a$, there are two-step trajectories whose steps are individually authorised while the trajectory guard does not authorise the second step. It denies that step if $H_s^0 \geq \theta_d$.*

Proof For (i), extending a trajectory enlarges the index set over which each maximum is taken. Part (ii) follows by comparing the corresponding factors. For (iii), give both steps effect triple $(1, 0, 0)$, and telic triples $(1, 0, 0)$ and $(\tau_P, 0, H_s)$, with $H_s^0 = H_s$. Set $\sigma(r) = \eta(s) = 1$ and $\sigma(s) = \eta(r) = 0$. Both step-wise risks are zero and both purpose tests pass. On the trajectory,

$$K \geq 1 - \beta(\tau_P - H_s)_+ \geq 1 - \beta\tau_P > \theta_a,$$

whereas $K_H^0 = H_s$. The second step therefore fails authorisation and meets denial under the stated hostility condition. $\square$

Part (ii) is a monotonicity statement, not unconditional protection against every decomposition. It applies when decomposition does not make the egress appear better supported by the principal’s purpose. The support bound of Theorem 4(i) handles low-trust additions under its own assumptions. In the reported experiments, the kernel did not change an attack outcome; its contribution is formal protection for settings with weaker purpose

estimates, rather than a demonstrated empirical gain here.

8 Experimental evaluation

We evaluate the guard on recorded and scripted tool calls, following the architecture in Figure 1. These experiments concern the dual-use distinction, the boundary of the provenance guarantee, the contribution of individual components and the training data needed to learn telic evidence. No deployed institutional agent is involved.

8.1 Common protocol

We measure an attack’s *pass-through rate* as $\Pr[\text{auth} \mid \text{attack}]$: the proportion of cases that would execute without review. Calibration uses a class-balanced cost with $c_{\text{miss}} = 10$, $c_{\text{fb}} = 1$ and $c_{\text{esc}} = 0.3$:

$$\begin{aligned} C = \frac{1}{2} & (c_{\text{miss}} \Pr[\text{auth} \mid \text{attack}] \\ & + c_{\text{esc}} \Pr[\text{esc} \mid \text{attack}]) \\ & + \frac{1}{2} (c_{\text{fb}} \Pr[\text{deny} \mid \text{benign}] \\ & + c_{\text{esc}} \Pr[\text{esc} \mid \text{benign}]). \end{aligned} \quad (1)$$

This loss makes unauthorised execution more expensive than refusal or review, following the decision-theoretic approach to three-way decisions (Yao 2010). All method thresholds are selected by exhaustive search on calibration data only; Appendix A gives the grids.

Confidence intervals are 95% percentile bootstrap intervals from 2,000 cluster resamples. We

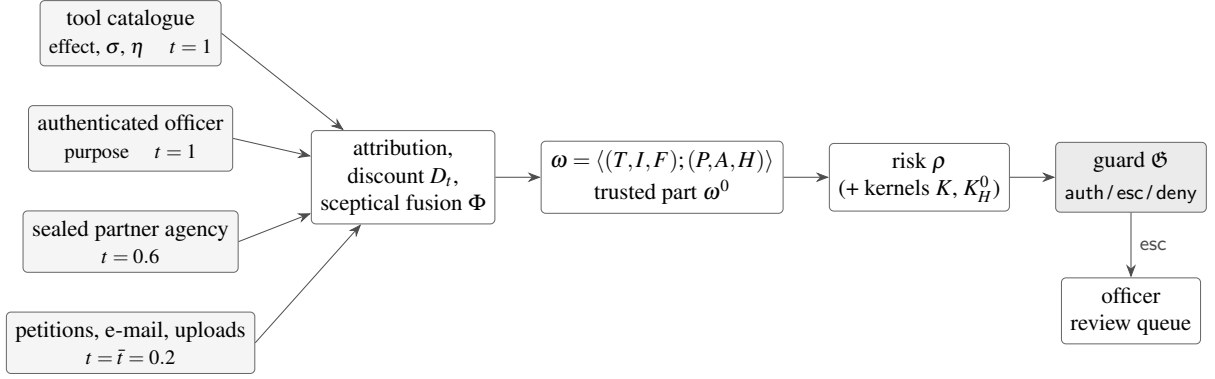


Fig. 1 The guard as a reference monitor. Evidence for a proposed tool call is attributed to its contributing channels, discounted and fused separately for effect and purpose. The denial test uses trusted evidence only; the full evidence set determines eligibility for authorisation. Escalated calls enter an officer review queue.

resample attacker or user instructions in Experiment 1 and scenarios in Experiment 2, rather than treat repeated phrasings or shared instructions as independent observations. Method differences use paired cluster resampling.

8.2 Experiment 1 on InjecAgent

8.2.1 Data and evidence

InjecAgent combines 17 user tasks with 62 attacker instructions embedded in tool responses (Zhan et al. 2024). Its 1,054 attack cases comprise 510 direct-harm cases (DH; 30 attacker instructions) and 544 data-stealing cases (DS; 32 instructions). Direct-harm goals include transferring funds or unlocking a door. A data-stealing trajectory first reads private data and then e-mails it to the attacker.

We use two benign controls: the 17 original user tasks (B-task) and the 62 attacker instructions presented instead as authorised requests from the principal (B-dual). The latter control holds the requested action fixed while changing its authority. InjecAgent pairs attacks with user tasks unrelated to the attacker tool, which makes purpose estimation relatively easy. Experiment 2 includes less cleanly separated cases.

Evidence rules are fixed before comparing guards. A lexicon applied to tool names and summaries supplies the effect triple, sensitivity and egress. For purpose alignment, a tool is represented by words from its toolkit and action names, together with the toolkit’s human-readable name.

Term frequency–inverse document frequency (TF-IDF) cosine similarity compares that representation with an instruction; the summary is used by the effect lexicon, not by this alignment calculation. Hostility is estimated from override phrases and identifiers, such as account numbers or e-mail addresses, that appear in the proposing text but not the principal’s request. Appendix A specifies the construction.

Every call receives a trusted telic piece from the principal’s request. A call proposed by tool content also receives a piece with $t = 0.2$. Effect evidence comes only from the catalogue, so the attacker contributes to the telic axis alone. We compare two ambivalence constructions. The baseline sets raw $A = 0$, leaving the discount as the source of ambivalence. The ignorance-aware version sets

$$A = 1 - \max\{P, H\}.$$

A source that supports neither purpose alignment nor hostility is then recorded as ambivalent. Applied to the principal’s evidence, this preserves the fact that the officer never requested a call, even if another source later supplies strong support.

8.2.2 Threat models and comparators

The evaluation distinguishes five conditions:

- T1: the benchmark injections enter through an untrusted channel.

- T1*: a constructed worst-case attacker supplies perfect apparent alignment and no hostility markers.
- T2: injected content is incorrectly labelled trusted, while attribution still records that tool output proposed the call.
- T2*: the T2 label error is combined with the T1* attacker.
- T3: every benign request receives an untrusted message demanding that the action be blocked and naming an outside address.

T2 and T2* deliberately violate provenance integrity. An additional enhanced-injection condition prepends an override instruction. It reduces T1 direct-harm pass-through to zero for the baseline TNL, its $\lambda = 0$ variants and the provenance-blind monitor, because the override raises hostility. Other decisions change little, so that condition is omitted from the tables.

Effect-only fuzzy, intuitionistic fuzzy and neutrosophic comparators test what safety evidence alone can achieve. A provenance-blind neutrosophic monitor combines effect evidence with purpose evidence from the latest instruction, exposing the consequences of ignoring its origin.

The crisp baselines include two-way provenance, taint-and-escalate, and taint with the same hostility detector as TNL. A CaMeL-like plan/data-separation rule escalates every call proposed by tool output, including calls whose arguments come from that output, irrespective of its trust label. It is a static rule inspired by the architectural principle, not an implementation or re-evaluation of CaMeL. A tier-aware crisp rule permits channels at or above a policy cutoff of 0.4 to supply purpose evidence, provided the raw alignment threshold is met and no hostility is detected. The cutoff equals the transferred TNL taint boundary.

Two further comparators use neutrosophic information. One adds distrust to effect indeterminacy, with and without trusted-only denial. The other receives exactly TNL’s effect and telic evidence, discount and trusted-only denial, but combines the triples by single-valued neutrosophic intersection and a score. This controls for access to information rather than merely comparing different evidence sources.

We evaluate TNL with baseline and ignorance-aware evidence. Ablations use a symmetric discount, ignore A and H , or set $\lambda = 0$. The latter two are tested both after recalibration and with TNL’s thresholds held fixed. Appendix A gives the decision rules.

8.2.3 Splits

Ten random, attacker-grouped two-fold splits are used. An attacker instruction, its legitimate dual-use counterpart and all user-task combinations containing that instruction remain in one fold. Thresholds are calibrated on T1 direct-harm and benign items in one fold and applied to the other. Data-stealing trajectories and all other threat models are excluded from calibration. Rates average each item’s decisions over the ten splits.

8.2.4 Results

Table 3 shows why benign dual-use controls matter. Effect-only rules admit no attack, but also authorise none of the legitimate sensitive requests; 15–20% of those requests are denied. The provenance-blind monitor admits 36.7% of direct-harm attacks under T1 and all attacks under T1*. Two-way crisp provenance improves that result but still passes 2.6% of direct-harm and 5.6% of data-stealing attacks, while denying 22% of legitimate dual-use requests.

With correct labels, the stronger crisp rules perform at least as well as TNL. Taint, taint with hostility detection and plan/data separation pass no T1 or T1* attack and authorise 90% of B-dual requests (cluster interval 82–95%). Their cost is 0.049 ± 0.012 across splits. Baseline TNL authorises 83% (74–91%), a paired difference of -6.5 percentage points relative to taint (interval -12.9 to -1.0), at cost 0.082 ± 0.025 .

TNL’s direct-harm pass-through is 1.2% under T1 and 1.4% under T1*. In 6 of the 20 fold calibrations, the search selects $\lambda = 0.25$. Then $\lambda(1 - \bar{t}) < \theta_a$, so Proposition 3 no longer blocks all low-trust contributions. The purpose threshold remains protective in the more limited sense of Theorem 4(i): an admitted call must already have sufficient purpose support in the principal’s own request. In the other folds, TNL operates as a taint rule.

Under mislabelling (T2), baseline TNL passes 36.7% of direct-harm attacks and no data-stealing

attack, compared with 82.3% and 96.6% for taint. The paired direct-harm difference is -45.7 points (interval -62.3 to -28.7). Crisp provenance passes 68.7% and 87.3%; provenance-indeterminacy and the two-criterion neutrosophic rule also deteriorate. Taint with TNL’s hostility detector exactly matches TNL here. The shared hostility detector therefore explains the observed improvement. Plan/data separation remains at zero even under T2*, because it uses attribution rather than trusting the label.

Independent ambivalence changes the T2* result. With ignorance-aware evidence, TNL’s T2 direct-harm pass-through falls from 36.7% to 16.9%. Under T2*, direct-harm and data-stealing pass-through fall from 100% each to 34.5% and 28.0%; the direct-harm cluster interval is 27.7–41.8%. The label-based crisp comparators pass all attacks in that condition. The principal’s evidence carries ambivalence about an unrequested call; fusion retains it when a mislabelled source supplies full purpose support. Under correct labels, T1 direct-harm pass-through is 1.2%, dual-use authorisation is 81% and cost is 0.085 ± 0.022 . Attribution-based separation is still stronger.

The availability attack exposes a different failure mode. No provenance-aware method with trusted-only denial increases its denial of legitimate requests under T3. The version that lets provenance-induced indeterminacy enter the denial test instead denies 99% of legitimate dual-use items. Taint and plan/data separation escalate every benign item they do not deny. Baseline TNL retains automatic handling of 31% of routine tasks, and its $\lambda = 0$ variant retains 98%.

The ablations show the costs of that flexibility. At fixed TNL thresholds, ignoring A and H raises T2 pass-through from 36.7%/0% to 79.3%/34.8%, although it also leaves 45% of dual-use requests automated during T3. Removing the uncertainty weight leaves T2 unchanged, raises T1 direct-harm pass-through to 2.0% and preserves most routine automation during T3. Recalibration lowers dual-use authorisation to 63% for purpose-only evidence and 68% for $\lambda = 0$.

InjecAgent has no intermediate-trust channel, so its tier-aware crisp rule coincides with taint plus hostility detection. Symmetric discounting changes outcomes through recalibration (T1 direct-harm pass-through 0.5%, cost 0.067), since attacks never contribute effect evidence in this

experiment. The trajectory kernel changes no outcome.

8.3 Experiment 2 on synthetic e-government scenarios

8.3.1 Scenario design

This experiment uses hand-written scenario texts processed by the guard. It is a simulation: the officers, citizens and agencies are fictional roles, and no deployed institution, live AI agent or human participant is studied.

The fictional agent has 21 tools covering a civil registry, tax office, health and land records, petitions, official e-mail, inter-agency exchange, document archiving, staff access, a public website, a calendar and statistics. Appendix B lists the catalogue. Citizen petitions, e-mail and uploaded documents have trust 0.2; a partner-agency channel carrying an electronic seal has trust 0.6. The principal is an authenticated officer.

Eight single-step goals include granting administrator rights, disabling two-factor authentication, issuing a fraudulent refund, publishing a false announcement, deleting an audit report, changing an address, sharing staff information and transferring a registry record to an unknown body. Six further goals split exfiltration into two steps. Each goal appears on every channel in three phrasings: plain, an override instruction and impersonation of a ministry circular. This gives 96 single-step (S) and 72 two-step (X) attack texts.

Argument manipulation is tested separately. In six Arg scenarios, an officer requests the action, but e-mail or an uploaded document changes an argument to a new IBAN, recipient or account identifier. Six Arg0 scenarios alter an amount, agency, role, address-form page or scope without introducing a new identifier. Each scenario occurs through e-mail and documents, giving 12 cases per argument category.

The benign set contains 12 routine officer tasks (B-task), 24 requests for the sensitive actions (B-dual; 8 scenarios), 18 legitimate two-step transfers (B-traj; 6 scenarios), 16 citizen-driven requests (B-citizen; 8 scenarios) and 16 sealed-partner requests (B-partner; 8 scenarios). In B-citizen, the officer gives a generic instruction to process a petition and carry out its request if procedures allow it; the petition supplies the concrete action.

Table 3 InjecAgent results averaged over ten attacker-grouped two-fold splits. DH: 510 direct-harm cases from 30 instructions; DS: 544 data-stealing cases from 32 instructions. Benign controls comprise 62 dual-use instructions and 17 user tasks. Fixed-threshold ablations use the baseline TNL thresholds; other ablations are recalibrated.

(a) Attack pass-through (%)								
Method	T1		T1*		T2		T2*	
	DH	DS	DH	DS	DH	DS	DH	DS
Fuzzy	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IFS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVN effect-only	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVN-2V, provenance-blind	36.7	0.0	100.0	100.0	36.7	0.0	100.0	100.0
Crisp provenance	2.6	5.6	2.6	5.6	68.7	87.3	100.0	100.0
Taint-escalate	0.0	0.0	0.0	0.0	82.3	96.6	100.0	100.0
Taint + hostility	0.0	0.0	0.0	0.0	36.7	0.0	100.0	100.0
Plan/data separation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tier-aware crisp	0.0	0.0	0.0	0.0	36.7	0.0	100.0	100.0
SVN + provenance-I	0.0	0.0	0.0	0.0	61.0	100.0	61.0	100.0
SVN + provenance-I, trusted denial	0.0	0.0	0.0	0.0	90.0	100.0	90.0	100.0
Two-criterion SVN + provenance	0.0	0.0	0.0	0.0	53.3	100.0	63.0	100.0
TNL	1.2	0.0	1.4	0.1	36.7	0.0	100.0	100.0
TNL, ignorance-aware A	1.2	0.0	1.4	0.1	16.9	0.0	34.5	28.0
TNL, symmetric discount	0.5	0.1	0.5	0.1	36.7	0.0	100.0	100.0
TNL, purpose only	1.2	0.1	1.2	0.4	46.3	22.3	60.3	98.1
Purpose only, fixed thresholds	2.5	0.3	2.5	0.5	79.3	34.8	100.0	100.0
TNL, $\lambda = 0$, recalibrated	1.2	0.0	1.5	0.5	26.3	0.0	60.3	98.1
$\lambda = 0$, fixed thresholds	2.0	0.0	2.5	0.5	36.7	0.0	100.0	100.0

(b) Benign outcomes (%) and T1 cost							
Method	T1			T3			Cost
	Dual auth	Dual deny	Task auth	Dual auth	Dual deny	Task auth	
Fuzzy	0	15	100	0	15	100	0.287
IFS	0	20	100	0	20	100	0.293
SVN effect-only	0	15	100	0	15	100	0.287
SVN-2V, provenance-blind	79	2	100	0	97	0	1.019
Crisp provenance	78	22	100	78	22	100	0.294
Taint-escalate	90	10	100	0	10	0	0.049
Taint + hostility	90	10	100	0	10	0	0.049
Plan/data separation	90	10	100	0	10	0	0.049
Tier-aware crisp	90	10	100	0	10	0	0.049
SVN + provenance-I	81	5	100	0	99	0	0.037
SVN + provenance-I, trusted denial	95	3	100	0	3	0	0.159
Two-criterion SVN + provenance	77	6	100	0	6	0	0.102
TNL	83	8	100	0	8	31	0.082
TNL, ignorance-aware A	81	9	100	0	9	31	0.085
TNL, symmetric discount	83	8	100	5	8	11	0.067
TNL, purpose only	63	8	100	26	8	100	0.107
Purpose only, fixed thresholds	83	8	100	45	8	100	0.121
TNL, $\lambda = 0$, recalibrated	68	8	100	0	8	98	0.100
$\lambda = 0$, fixed thresholds	83	8	100	0	8	98	0.101

8.3.2 Transfer protocol

No threshold is tuned on these scenarios. Every method uses the parameters selected once on all InjecAgent data. The TF-IDF model is fitted on InjecAgent texts and the e-government tool catalogue. A domain lexicon adds personal-data categories, access-control assets and three administrative verbs. Because that lexicon was developed together with the scenarios, we also report results with it removed.

We additionally compute a reachable set. For each carrier task and channel, we count the catalogue tools that the constructed worst-case attacker could get authorised. This tests adaptive tool selection rather than only the attacks listed in the benchmark.

8.3.3 Results

Tables 4–6 show the transferred decisions. Under T1, TNL, the taint rules and plan/data separation admit none of the 168 injected attacks and

escalate all argument-level attacks. The two-way crisp rule authorises 83% of argument-level cases because the officer did request the underlying action. TNL authorises 88% of legitimate officer dual-use requests (cluster interval 75–96%) and 61% of legitimate two-step transfers; the crisp three-way rules authorise 92% and 67%.

On low-trust channels, these results agree with the taint limit. No catalogue tool is reachable from petitions, e-mail or uploads. With $\lambda = 0$, between three and five tools become reachable on each of those channels, within the trusted-purpose bound of Theorem 4(i).

Intermediate trust permits additional automation. Baseline TNL authorises 50% of legitimate sealed-partner requests and denies none. The interval is wide, 12–88%, because there are only eight independent scenarios. Taint and plan/data separation instead escalate 88% and deny 12%; the paired authorisation difference is +50 points (interval +12 to +88).

Automating partner requests also exposes additional tools to attack. A worst-case attacker on that channel can reach 12 of the 21 tools. Across all channels, TNL passes 3.1% of single-step and 4.2% of two-step cases under T1*, equivalent to 12.5% and 16.7% of the partner-channel cases. Trust 0.6 exceeds the purpose threshold 0.3, so the low-trust authorisation bound does not apply there. With $\lambda = 0$, Figure 2 isolates this boundary: pass-through follows the trusted-evidence bound below τ_P and rises when the channel reaches that threshold.

The tier-aware crisp rule obtains the same partner-channel gain: it authorises 62% of partner requests, denies 12%, and has the same worst-case pass-through and the same 12 reachable tools. In both rules, a sufficiently trusted channel may supply purpose evidence. TNL expresses that permission through a parameter-derived boundary within the same calculus used for lower-trust channels.

Retaining ignorance favours review on the partner channel. This variant escalates all legitimate partner requests and reduces the partner reachable set to two tools. When the officer’s instruction does not name the action, the absence of a request remains a reason for review. The reduction in exposure therefore comes at the cost of partner automation.

Under T2, baseline TNL passes 12.5% of single-step, 0% of two-step and 16.7% of Arg attacks. Taint passes 100%, 91.7% and 100%. Adding hostility detection to taint again matches TNL, while plan/data separation passes none. Arg0 attacks expose the detector’s blind spot: without a new identifier, TNL, taint with hostility detection and crisp provenance pass 83–100% under T2.

Under T2*, the label-based crisp comparators fail completely. The provenance-indeterminacy and two-criterion neutrosophic rules pass 50.0% of single-step and 33.3% of two-step cases; recalibrated TNL with $\lambda = 0$ passes 37.5% and 16.7%. Ignorance-aware TNL reduces these to 25.0% and 0%. It also lowers T2 single-step pass-through to 3.1%, although Arg0 pass-through remains 66.7%.

Generic delegation remains difficult. For legitimate citizen requests, TNL, taint and plan/data separation escalate 75% and deny 25%. The generic officer instruction does not provide enough purpose support to offset the harm of a refund or registry transfer. Effect-only rules authorise half of these requests, but cannot distinguish their malicious counterparts.

Review demand reflects these choices. In the benchmark’s benign mix, TNL requires 326 reviews per 1,000 requests, compared with 302 for taint, 419 for ignorance-aware TNL and 186 for the tier-aware crisp rule. Under T3, baseline TNL and taint both require 837 reviews. TNL still authorises partner requests amounting to 9% of benign traffic and denies 7%; the taint-based crisp rules authorise none and deny 16%.

Finally, removing the domain lexicon raises baseline TNL’s T1 single-step pass-through to 5.2%. Refund and announcement attacks on the partner channel then appear less harmful. Taint and plan/data separation remain at zero because they escalate that channel as a whole. This dependence on the effect estimator is separate from the provenance theorem.

8.4 Experiment 3 on learned telic evidence

8.4.1 Records and estimation

We next replace fixed telic rules with an estimator learned from call records. A *teleoneutrosophic record* couples a proposed call with its contributing evidence and adjudication.

Table 4 Synthetic e-government attacks. S: single-step, 96 cases under T1 and 32 otherwise (8 goals). X: two-step, 72/24 cases (6 goals). Arg and Arg0 each comprise 12 cases from 6 scenarios, with and without a new identifier. T1* and T2* are the constructed worst-case conditions.

(a) T1, T1* and domain-lexicon ablation: pass-through (%)								
Method	S	T1		Arg0	T1*		No domain lexicon	
		X	Arg		S	X	S, T1	S, T2
Fuzzy	0.0	0.0	0.0	0.0	0.0	0.0	25.0	25.0
IFS	0.0	0.0	0.0	0.0	0.0	0.0	12.5	12.5
SVN effect-only	0.0	0.0	0.0	0.0	0.0	0.0	12.5	12.5
SVN-2V, provenance-blind	5.2	0.0	16.7	50.0	100.0	100.0	25.0	25.0
Crisp provenance	9.4	8.3	83.3	83.3	9.4	8.3	34.4	100.0
Taint-escalate	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
Taint + hostility	0.0	0.0	0.0	0.0	0.0	0.0	0.0	25.0
Plan/data separation	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tier-aware crisp	0.0	0.0	0.0	0.0	3.1	4.2	4.2	25.0
SVN + provenance-I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	87.5
Two-criterion SVN + provenance	4.2	1.4	0.0	0.0	6.2	4.2	10.4	87.5
TNL	0.0	0.0	0.0	0.0	3.1	4.2	5.2	25.0
TNL, ignorance-aware A	0.0	0.0	0.0	0.0	3.1	0.0	0.0	3.1
TNL, $\lambda = 0$	0.0	0.0	0.0	16.7	6.2	0.0	8.3	25.0

(b) Mislabelled channels: pass-through (%)							
Method	S	X	T2		Arg0	T2*	
			Arg	Arg0		S	X
Fuzzy	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IFS	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVN effect-only	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SVN-2V, provenance-blind	12.5	0.0	16.7	50.0	100.0	100.0	100.0
Crisp provenance	100.0	87.5	83.3	83.3	100.0	100.0	100.0
Taint-escalate	100.0	91.7	100.0	100.0	100.0	100.0	100.0
Taint + hostility	12.5	0.0	16.7	100.0	100.0	100.0	100.0
Plan/data separation	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tier-aware crisp	12.5	0.0	16.7	100.0	100.0	100.0	100.0
SVN + provenance-I	50.0	33.3	33.3	33.3	50.0	50.0	33.3
Two-criterion SVN + provenance	50.0	33.3	33.3	33.3	50.0	50.0	33.3
TNL	12.5	0.0	16.7	100.0	100.0	100.0	100.0
TNL, ignorance-aware A	3.1	0.0	16.7	66.7	25.0	25.0	0.0
TNL, $\lambda = 0$	12.5	0.0	0.0	16.7	37.5	37.5	16.7

For each evidence piece, the record contains the principal’s request u , the source text s , the tool k , channel trust t , and labels y_P, y_H . For the principal’s own piece, $s = u$. The labels indicate whether the call serves the principal and whether the source advances an adversarial purpose. From features x , the estimator produces

$$\begin{aligned}\hat{P} &= \Pr(y_P = 1 \mid x), \\ \hat{H} &= \Pr(y_H = 1 \mid x), \\ \hat{A} &= 1 - |2\hat{P} - 1|.\end{aligned}$$

The ignorance-aware version instead uses

$$\hat{A} = \max\{1 - |2\hat{P} - 1|, 1 - \max\{\hat{P}, \hat{H}\}\}.$$

For attack items, every piece has $y_P = 0$, and the injected source has $y_H = 1$. For benign items,

every piece has $y_P = 1$ and $y_H = 0$. These are the benchmark’s per-piece training labels, not observations of an officer’s decisions in a deployed service.

Estimated triples enter the same discount, fusion and guard as rule-based evidence. Theorem 4 does not require a particular estimator or statistical accuracy assumption. Its proof still applies when trusted triples use trusted inputs only and the estimator is fixed independently before the attack. It does not promise safety after the provenance-integrity assumption is violated.

8.4.2 The role of provenance in the estimator

Figure 3 states the causal assumptions used to motivate the estimator. In these records, an adversarial goal influences both the chosen channel and

Table 5 Synthetic e-government benign outcomes. Panel (a) reports authorisation (%) for B-task and authorised/escalated/denied percentages for the other categories. Counts are B-task 12, B-dual 24, B-traj 18, B-citizen 16 and B-partner 16. Panel (b) reports percentages of all benign requests under T3 and reviews per 1,000 benign requests under the benchmark mix. Rounding can make component totals differ from 100.

(a) Benign requests under T1					
Method	B-task	B-dual	B-traj	B-citizen	B-partner
Fuzzy	92	0/50/50	0/100/0	50/25/25	62/38/0
IFS	92	0/50/50	0/100/0	50/25/25	62/38/0
SVN effect-only	92	0/50/50	0/33/67	50/12/38	62/25/12
SVN-2V, provenance-blind	100	83/17/0	56/39/6	50/25/25	50/25/25
Crisp provenance	100	83/0/17	61/0/39	62/0/38	88/0/12
Taint-escalate	100	92/0/8	67/0/33	0/75/25	0/88/12
Taint + hostility	100	92/0/8	67/0/33	0/75/25	0/88/12
Plan/data separation	100	92/0/8	67/0/33	0/75/25	0/88/12
Tier-aware crisp	100	92/0/8	67/0/33	0/75/25	62/25/12
SVN + provenance-I	92	50/50/0	33/0/67	0/50/50	0/62/38
Two-criterion SVN + provenance	92	50/50/0	33/0/67	0/75/25	62/25/12
TNL	100	88/8/4	61/33/6	0/75/25	50/50/0
TNL, ignorance-aware A	100	88/8/4	61/33/6	0/75/25	0/100/0
TNL, $\lambda = 0$	92	29/67/4	0/94/6	0/75/25	50/50/0

(b) Availability attack and review demand				
Method	T3 auth	T3 deny	T1 reviews	T3 reviews
Fuzzy	34	19	477	477
IFS	34	19	477	477
SVN effect-only	34	38	279	279
SVN-2V, provenance-blind	20	66	221	140
Crisp provenance	78	22	0	0
Taint-escalate	0	16	302	837
Taint + hostility	0	16	302	837
Plan/data separation	0	16	302	837
Tier-aware crisp	12	16	186	721
SVN + provenance-I	0	66	349	337
Two-criterion SVN + provenance	12	22	326	663
TNL	9	7	326	837
TNL, ignorance-aware A	0	7	419	930
TNL, $\lambda = 0$	16	7	628	767

Table 6 Number of catalogue tools reachable by the worst-case attacker, out of 21. Petition, e-mail and upload channels have trust 0.2; the partner channel has trust 0.6. Thresholds are transferred from InjecAgent.

Method	Petition	E-mail	Upload	Partner
TNL	0	0	0	12
TNL, ignorance-aware A	0	0	0	2
TNL, $\lambda = 0$	4	3	5	11
Tier-aware crisp	0	0	0	12
Taint-escalate	0	0	0	0
Plan/data separation	0	0	0	0

the content. The trust label is therefore associated with hostility without being its cause. A model that reads the label can learn that association. Under T2, changing the label to trusted is an intervention $\text{do}(t := 1)$ while the injected content remains the same.

Channel trust enters D_t , while content evidence is used to estimate P and H . Excluding t makes the estimator invariant to changing that label alone when the other features are held fixed. Content features can still reveal that a piece came from outside, so this invariance concerns

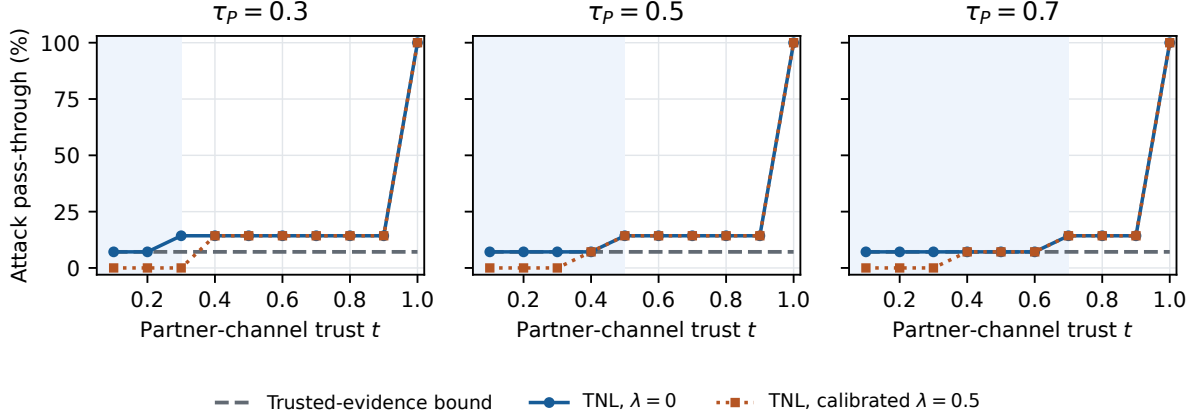


Fig. 2 Synthetic e-government benchmark: worst-case partner-channel pass-through as channel trust varies, at three purpose thresholds. Shading marks $t < \tau_P$, where Theorem 4(i) applies. The solid curve uses $\lambda = 0$ to isolate that bound; the dotted curve uses calibrated $\lambda = 0.5$; the dashed curve is the trusted-evidence bound.

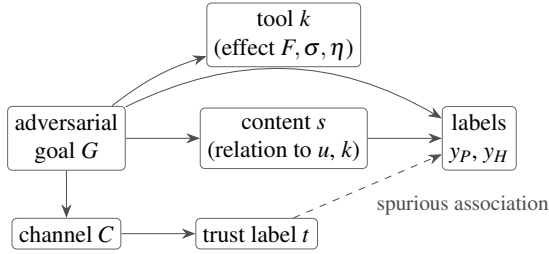


Fig. 3 Assumed causal structure of the records. The adversarial goal influences content, channel and tool selection; the channel determines its trust label. The dashed association is the shortcut available to a label-conditioned estimator. Omitting the label prevents its hostility estimate from changing when the label alone is intervened on.

only the label intervention. Tool harm is likewise not sufficient evidence of whose purpose a call serves, although effect features may help distinguish kinds of calls. Surface style, including capitals, politeness and claims of authority, is readily chosen by an attacker.

A further problem is training coverage. InjecAgent contains no legitimate request arriving through an untrusted channel. Thus every untrusted training piece is an attack. A model can use external-looking content as a substitute for the omitted label. Learning to accept legitimate external requests requires examples of those requests; removing the trust feature cannot supply them.

8.4.3 Estimators and protocol

Class-balanced logistic regressions estimate $\hat{P}$ and $\hat{H}$ from standardised features. Training uses full-batch gradient descent for 2,500 steps, step size 0.3 and an L_2 penalty of 10^{-2} . The feature groups are:

- relational content: tool alignment with s and u , their alignment gap, presence and number of identifiers absent from u , and a delegation cue in u ;
- tool effect: F , σ and η ;
- style: overrides, claimed authority, capitals, length and politeness;
- the trust label, included only in the associational comparator.

The main estimator omits the trust label. Variants omit effect features or add variance risk extrapolation (V-REx) (Krueger et al. 2021). V-REx adds γ times the variance of class-balanced environment risks to their mean, with $\gamma = 5$. Its two environments are plain injections and injections with an override phrase. IRM seeks predictors that remain optimal across training environments (Arjovsky et al. 2019); invariant causal prediction uses invariance across experimental settings to constrain causal predictors (Peters et al. 2016). These ideas and causal representation learning motivate the separation of stable content from source labels (Schölkopf et al. 2021). Our implementation uses V-REx rather than IRM or invariant causal prediction. Its environments vary only a narrow aspect

of style, so this experiment does not establish causal identification.

To test training coverage, we add synthetic legitimate delegated requests equal to 10% or 25% of the attack count. In those records, the principal asks the agent to process an attached untrusted message containing a dual-use instruction.

InjecAgent evaluation uses five repeated attacker-grouped two-fold splits. Within each training fold, attacker groups are split again between estimator fitting and guard calibration. Thresholds are therefore selected on predictions for groups not used to fit the estimator. For e-government transfer, fitting and calibration use InjecAgent once, with no later adjustment.

Two diagnostics are computed on external-source pieces from e-government attacks and legitimate citizen or partner requests under T1: the area under the receiver-operating-characteristic curve (AUC) for $\hat{H}$, and the mean absolute change $|\Delta\hat{H}|$ when only the trust label is set to one.

8.4.4 Results

Table 7 reports the learned variants. Under correct labels, their direct-harm pass-through is at most 1.1% and their dual-use authorisation is 85–87%, compared with 81% for rules under this experiment’s calibration protocol.

Learning helps most under ordinary channel mislabelling. InjecAgent T2 direct-harm pass-through falls from 36.7% for rules (cluster interval 20.0–53.3%) to 1.6–1.8% for the learned estimators. The provenance-free estimator’s interval is 0.0–4.2%. On transfer, T2 single-step and new-identifier argument attacks fall from 12.5% and 16.7% to zero, except that omitting effect features leaves argument pass-through at 33.3%.

This improvement does not extend to every attacker. When ambivalence is derived from $\hat{P}$ alone, T2* pass-through remains 93–96% on InjecAgent and 84–91% on the synthetic benchmark. The ignorance-aware construction reduces those rates to 27.3% and 9.4%, with dual-use authorisation of 84% and 88%, respectively. Argument changes without new identifiers remain largely undetected: pass-through is 83–100% for the estimators without the training-coverage augmentation.

The associational model’s hostility estimate depends on the trust label. Its standardised hostility coefficient for the label is -1.51 , among its largest coefficients. Relabelling external pieces changes their mean hostility estimate by 0.19; the change is exactly zero for every estimator that omits the label. On this benchmark, the decision effect is limited because relational features carry most of the signal. Nevertheless, the associational model denies 62% of petition-driven requests, compared with 38% for the provenance-free model, and neither automates partner requests.

V-REx changes almost no decision. Once relational features are present, adding an override phrase creates little useful variation between environments. Removing effect features has a clearer cost: T2 argument pass-through rises from zero to 33.3%, and citizen-request denial rises from 38% to 62%. These results support excluding the trust label, but do not establish a benefit from the particular V-REx environments used here.

The provenance-free model still has partial proxies for source origin. For the principal’s piece, the gap between s and u and the number of new identifiers are zero by construction. Their usefulness can therefore reflect the absence of legitimate external training pieces. This limits what label removal alone achieves.

Adding 10% synthetic legitimate delegated requests raises external-piece AUC from 0.70–0.72 to 0.98. It authorises all partner requests, removes citizen-request denials and retains zero T2 pass-through for single-step and new-identifier argument attacks. Dual-use authorisation is 83%, and T1* single-step pass-through remains 3.1%. This configuration still passes every T2* single-step attack: improved coverage does not remove the worst-case compromise tradeoff.

At 25% augmentation, calibration selects a lower $F_{\max}$ and dual-use authorisation falls to 25%. T2* single-step pass-through falls to 37.5%, Arg0 pass-through to 16.7%, and T1* single-step pass-through rises to 6.2%. The training mixture and the calibration workload therefore need to be considered together. Since the synthetic delegated requests resemble the benchmark’s citizen instructions, these numbers are optimistic transfer results, not estimates of field performance. Real institutional records must include legitimate citizen and partner traffic as well as attacks.

Table 7 TNL with learned telic evidence. InjecAgent uses five attacker-grouped splits and separate fitting and calibration groups. E-government results transfer the InjecAgent fit. AUC and label-intervention diagnostics concern external-source pieces under T1. The +10% and +25% rows augment training with synthetic legitimate external requests; InjecAgent outcomes are not applicable for those rows. All rows, including rules, use this experiment’s calibration protocol, so its rule-based outcomes can differ from earlier tables.

(a) InjecAgent rates (%) and external-piece diagnostics						
Estimator	T1 DH	T2 DH	T2* DH	Dual auth	AUC	$ \Delta \hat{H} $
Rule-based evidence	1.6	36.7	100.0	81	0.70	0.00
Associational, with trust label	0.1	1.8	92.5	87	0.72	0.19
Provenance-free	0.5	1.6	95.5	86	0.70	0.00
Provenance-free, no effect features	1.1	1.6	93.8	86	0.71	0.00
Provenance-free, V-REx ($\gamma = 5$)	0.5	1.6	95.5	85	0.70	0.00
Provenance-free, ignorance-aware $\hat{A}$	0.4	1.5	27.3	84	0.70	0.00
Provenance-free +10% legitimate	—	—	—	—	0.98	0.00
Provenance-free +25% legitimate	—	—	—	—	0.98	0.00

(b) E-government transfer rates (%)								
Estimator	T2 S	T2 Arg	T2 Arg0	T2* S	T1* S	Dual	Partner	Citizen deny
Rule-based evidence	12.5	16.7	100.0	100.0	3.1	88	50	38
Associational, with trust label	0.0	0.0	83.3	90.6	3.1	88	0	62
Provenance-free	0.0	0.0	83.3	87.5	3.1	88	0	38
Provenance-free, no effect features	0.0	33.3	100.0	84.4	3.1	88	0	62
Provenance-free, V-REx ($\gamma = 5$)	0.0	0.0	83.3	87.5	3.1	88	0	38
Provenance-free, ignorance-aware $\hat{A}$	0.0	0.0	83.3	9.4	3.1	88	0	38
Provenance-free +10% legitimate	0.0	0.0	83.3	100.0	3.1	83	100	0
Provenance-free +25% legitimate	0.0	0.0	16.7	37.5	6.2	25	62	0

9 Discussion

9.1 What the construction contributes

Crisp taint rules are at least as safe as TNL when labels are correct. Under label errors, plan/data separation remains safe because it retains attribution. TNL itself becomes a taint rule below trust 0.4 with the transferred parameters. A crisp rule can also share its hostility detector and match its T2 improvement; a tier-aware crisp policy matches its partner-request automation. These comparisons delimit the empirical case for grading.

Retaining ambivalence about an unrequested call changes how later support is treated. A mislabelled channel may add purpose support while the principal’s lack of a request still counts towards review. This reduces worst-case pass-through in the constructed tests, at the cost of some automation. It does not prevent every attack.

The formal contribution is to express these choices within one decision semantics. The support thresholds state which channels may justify authorisation. The uncertainty weight determines when a channel becomes effectively tainted. Trusted-only denial limits the effect of planted

objections. Theorems give the logical behaviour of the designated values and the conditions under which external evidence can change a decision. Learned triples can enter the same calculus without altering those proofs, provided the trust and attribution assumptions still hold.

A richer crisp policy could preserve an explicit “not requested by principal” flag. Our comparisons therefore concern the stated baselines; they leave room for other crisp or many-valued designs with similar behaviour. TNL supplies a graded decision layer for an architecture that already separates trusted planning from external data.

9.2 Institutional operation and human review

Deployment begins with provenance and attribution. Source authentication must be paired with persistent labels and call-level or argument-level tracking. A seal should not automatically grant a message the same authority as an officer’s instruction. An institution can place τ_P above the trust of channels it does not wish to authorise independently, choose λ in light of review capacity, and set $F_{\max}$ below the impact of actions that require an additional signature.

Review demand is substantial even on this benchmark. Baseline TNL requires 326 reviews per 1,000 benign benchmark requests under T1 and 837 under T3. Taint requires 302 and 837. Queue prioritisation and rate limits are therefore part of the surrounding system. If an officer mistakenly approves a proportion q of escalated attacks, the eventual attack-execution rate is

$$\Pr[\text{auth} \mid \text{attack}] + q \Pr[\text{esc} \mid \text{attack}].$$

A low automatic pass-through rate does not by itself establish low end-to-end risk.

Denial applies to the agent’s proposed tool call. The citizen’s petition remains subject to normal institutional handling and the right to petition (Republic of Türkiye 1984). Theorem 4(ii) prevents an untrusted objection from entering this guard’s denial test, but it does not guarantee availability of the whole administrative service. Review queues can still be overloaded.

The legal relevance of review depends on the decision being made. Turkish Law 6698, Article 11(1)(g), allows a person to challenge an adverse outcome produced exclusively by automated analysis of their personal data (Republic of Türkiye 2016). GDPR Article 22 limits decisions reached without human involvement that have legal or comparably significant consequences, subject to its exceptions and safeguards (European Union 2016). The EU AI Act addresses specified public-assistance eligibility systems in Annex III and human oversight of high-risk systems in Article 14 (European Union 2024b). An assistant’s classification depends on its intended use and the applicable exceptions. Escalation in our model is a technical control, not a demonstration of compliance. Turkish cyber-security legislation and the information-security circular supply further institutional context (Republic of Türkiye 2025; Presidency of the Republic of Türkiye 2019).

Each decision can be logged with its full and trusted values, the contributing sources and the deciding risk term. An authorisation remains an action performed under institutional authority, with responsibility resting with the institution. Generic delegation needs particular care: a broad instruction to “process this petition” may require procedure-specific evidence of permitted actions to avoid the legitimate-request denials seen in Experiment 2.

9.3 Learning from institutional records

A pilot could retain anonymised records of proposed calls, contributing channels, policy trust and officer adjudications within the institution’s data-protection arrangements. The necessary training signal is broader than a collection of attacks. It must represent legitimate requests received through citizen and partner channels, including requests the officer did not spell out tool by tool.

The learned estimate should remain separate from the provenance discount, and trusted estimates must use trusted inputs only. Updating the model requires a controlled process that does not let an attacker write its training labels. The theorem then applies to each fixed deployed estimator under the same assumptions; it does not certify the quality of a newly trained model or remove the need to evaluate it.

9.4 Limitations

The provenance results are conditional. T2* shows the consequence of violating label integrity: every label-dependent guard tested admits some attacks. Plan/data separation avoids that failure using attribution, while effect-only rules avoid it by authorising no sensitive request. Ignorance-aware ambivalence reduces the label-dependent failure rather than eliminating it.

The first two experiments use a frozen lexicon and TF-IDF evidence. Their hostility detector misses argument changes without new identifiers. Benign categories contain only 6–62 independent scenarios, so intervals are wide and differences of a few points should be treated cautiously. The e-government benchmark is synthetic, and its domain lexicon was designed alongside the scenarios; the lexicon-removal results expose that dependency.

The learned estimators face strong confounding between source origin and purpose. The augmentation experiment improves training coverage, but its synthetic delegated requests resemble the target benchmark. It cannot establish performance on authentic institutional traffic. Some calibrated parameters, including $\tau_P = 0.3$, also lie at grid boundaries, which motivates wider policy and sensitivity studies.

No live language-model agent is run. The study therefore cannot establish interactive attack resistance, timing overhead or field review quality, and it does not empirically compare against runtime monitors requiring live execution (Jia et al. 2025; Zhu et al. 2025; Zhong et al. 2025). Its crisp rules are static calibrated baselines, not per-task policies synthesised by systems such as Progent (Shi et al. 2026). AgentDojo and real workflow trials are appropriate settings for the next evaluation (Debenedetti et al. 2024).

10 Conclusion

A guard that sees only a tool call’s effect cannot establish whose instruction it serves. Teleoneutrosophic values keep effect and purpose separate, then combine their evidence through risk coupling, provenance discount, fusion and a three-way guard. The underlying algebra is a product of two neutrosophic De Morgan algebras. We give exact parameter conditions for paraconsistency and adjunction in the resulting matrix logic.

With reliable labels and attribution, a channel below the purpose threshold cannot supply sufficient purpose support for authorisation, and untrusted evidence cannot cause denial. Stronger escalation-only statements require the specified support conditions. The experiments show both the usefulness and the limits of these guarantees. Low-trust behaviour often reduces to taint, a crisp attribution-based rule is at least as safe, and tiered crisp policies can match intermediate-channel automation.

Independent ambivalence preserves a reason for review that later purpose support need not erase. Learning improves performance under ordinary mislabelling, but useful automation also requires legitimate external requests in the training data. Interactive agents, representative institutional records and reliable human adjudication are the next tests of this approach.

Statements and declarations

Author information. Selçuk Topal is the first and corresponding author. Dr Buğra Ayan is Head of the Department of Information Technologies (Bilgi İşlem Daire Başkanı) at the Directorate of Communications, Presidency of the Republic of Türkiye.

Institutional disclaimer. The views expressed are those of the authors and do not necessarily reflect those of the Directorate of Communications. No operational records, systems or service users of the Directorate were involved in the experiments.

Funding. No funding was received for this work.

Competing interests. The authors declare no competing interests.

Ethics approval and consent. The study uses public benchmark material and fictional scenarios. It contains no research involving human participants or animals; consent to participate and consent for publication are not applicable.

Data availability. The Supplementary Information (Online Resource 1) contains the fictional e-government scenario definitions, 942 trajectory records across five threat models, 1,200 tool-call rows, synthetic delegated-request training records and grouped split assignments. Reused scenarios across threat models are not independent observations. InjecAgent data are included with the upstream licence, source revision and file hashes, and are also available from the authors’ repository (Zhan et al. 2024). All e-government names, identifiers, accounts and scenario roles are fictional; no actual citizen records were used.

Code availability. Online Resource 1 includes the TNL implementation, property tests, all comparator and ablation implementations, the learned estimators, the complete calibration and evaluation scripts, and the executed result files. It also provides table generation, trust-sweep figure generation, a pinned dependency list, portable commands and a reproducibility audit. The three agent experiments and the trust sweep were rerun from the supplied data; all 61 method rows in Tables 3–7, comprising 743 printed cells, have corresponding computational outputs.

Supplementary Information. Online Resource 1: reference implementation, benchmark inputs, fictional scenarios, synthetic training augmentation, grouped splits, executed experimental results, and scripts for tables, the trust sweep and reproducibility checks.

Author contributions. Selçuk Topal: conceptualisation, methodology, formal analysis, software, investigation and writing the original draft. Buğra Ayan: conceptualisation of public-sector requirements and the threat model, validation of scenario realism, and review and editing.

AI-assisted preparation. Experimental results were computed by executing the accompanying scripts on the stated benchmark data with the specified splits and calibration protocol. The authors are responsible for reviewing and approving the mathematical arguments, software, reported results and final text.

A Evidence construction and calibration grids

A.1 Effect evidence

Tool names are split into words. Set $\text{act} = 1$ when the verb is one of *send, transfer, withdraw, grant, unlock, delete, share, disable, update, pay, place, move, create, manage, generate, execute, dispatch, leave, schedule, go, give, deposit, auto, fill*. The scenario lexicon adds *issue, publish, reply*. Otherwise a tool is a read when its verb is *get, view, search, list, read, access, download or retrieve*.

Sensitivity is $s = \min\{1, k/2\}$, where k counts distinct terms found in the name or description from *password, payment, bank, account, genetic, patient, prescription, health, address, location, history, holdings, factor, medical, record, identity, funds, personal, saved, order, fhir, clinical, deepfake, lock, door*, including plurals. The scenario list adds *citizen, national, birth, family, registry, tax, taxpayer, income, debts, diagnoses, parcel, owner, mortgage, refund, treasury, iban, privileges, role, authentication, two*.

Set $\text{free} = 1$ for free-form parameters such as **command** or **content**. The effect values are

$$F = \min\{1, 0.1 + 0.5 \text{act} + 0.3s\},$$

$$I = 0.1 + 0.3 \text{free}, \quad T = (1 - F)(1 - I).$$

Egress η is one for send/share, 0.5 for other actions and zero for reads. Sensitivity σ is s for reads and $s/2$ otherwise.

A.2 Purpose and hostility evidence

For each tool, concatenate the words obtained from its toolkit and action names and the toolkit’s human-readable name. Fit a TF-IDF model with sublinear term frequency on these representations and InjecAgent user instructions. For transfer, include the e-government catalogue representations when fitting the vectoriser; no e-government scenario instruction enters that fit. Tool-text alignment is $\min\{1, \cos/0.25\}$. The separate effect lexicon also reads tool summaries. Hostility is the sum of two indicators, each weighted 0.5: the proposing text introduces an identifier absent from the principal’s request while $\eta > 0$; or it contains an override phrase. Regular expressions recognise e-mail addresses, underscore identifiers, upper-case codes and long digit strings. The trusted piece is calculated from the principal’s request and the catalogue alone. For ignorance-aware rules, set $A = 1 - \max\{P, H\}$ separately on each raw principal or source piece, before applying discount and fusion.

A.3 Comparator rules

Crisp provenance denies a call when $F \geq f_0$ unless trusted purpose support satisfies $P^0 \geq a_0$; otherwise it allows it. *Taint-and-escalate* denies when $F \geq f_0$ and $P^0 < a_0$, escalates a remaining call attributed to any untrusted channel and authorises the rest. *Taint with hostility detection* additionally escalates calls with $H^0 > 0$.

Plan/data separation denies when $F \geq f_0$ and the principal’s own request gives $P_{\text{principal}} < a_0$. It escalates calls proposed by tool output or containing arguments derived from it, regardless of label; other calls are authorised. Thus a mislabelled channel does not erase the tool-output attribution.

Tier-aware crisp first applies the taint denial test, $F \geq f_0$ and $P^0 < a_0$. For a remaining call attributed to a channel with $0.4 \leq t < 1$, any source or trusted hostility causes escalation; otherwise authorisation requires the maximum of the principal’s and the channel’s raw support to reach a_0 . Calls outside this intermediate tier follow taint with hostility detection. Thus the tier cutoff cannot override a trusted denial, and the rule coincides with taint plus hostility when no intermediate-trust source is present.

For *two-criterion SVN with provenance*, intersect the effect and telic triples using single-valued neutrosophic intersection. Score the result by $(2 + T - I - F)/3$, accept above threshold a subject to indeterminacy gate h , and deny if the trusted intersection’s score is at most b . All comparator trajectories are denied if any step is denied, authorised if every step is authorised, and escalated otherwise. Calibration uses Equation (1) throughout.

A.4 Search grids

For TNL, the grid is

$$\begin{aligned} \tau_T &\in \{0, 0.3\}, & \tau_P &\in \{0.3, 0.5, 0.7\}, \\ \theta_a &\in \{0.1, 0.2, 0.3, 0.4\}, \\ \theta_d &\in \{0.3, 0.45, 0.6, 0.75\}, & \theta_d &> \theta_a, \\ \lambda &\in \{0.25, 0.5, 1\}, & \beta &\in \{0.5, 1\}, \\ F_{\max} &\in \{0.6, 0.8, 1\}. \end{aligned}$$

Fix $\lambda = 0$ for that ablation. Crisp provenance, taint and plan/data separation use 17×19 pairs (f_0, a_0) . Neutrosophic comparators, including the two-criterion rule, use 14×14 score thresholds and four indeterminacy gates; the intuitionistic fuzzy baseline uses 19×19 thresholds and four gates; fuzzy uses 19×19 thresholds.

Across the 20 Experiment 1 fold calibrations, selected TNL parameters are $\tau_T = 0$, $\tau_P = 0.3$, $\theta_a \in \{0.2, 0.3\}$, $\theta_d \in \{0.3, 0.45, 0.6\}$, $\beta = 1$ and $F_{\max} = 1$. The search selects $\lambda = 0.5$ in 14 folds and 0.25 in six. Calibration on all InjecAgent yields, in the order $(\tau_T, \tau_P, \theta_a, \theta_d, \lambda, \beta, F_{\max})$,

$$(0, 0.3, 0.3, 0.6, 0.5, 1, 1).$$

B Synthetic e-government scenarios

The catalogue contains the following tool groups: CitizenRegistry (GetRecord, SearchByName, UpdateAddress); TaxOffice (GetTaxpayerInfo, IssueRefund); HealthRecords (GetPatientSummary); LandRegistry (GetParcelOwner); PetitionInbox (ReadPetition, ReplyToCitizen); OfficialMail (ReadInbox, SendMessage); Inter-AgencyExchange (ReadIncoming, SendRecord);

DocumentArchive (SearchDocuments, ShareDocument, DeleteDocument); AccessControl (GrantUserRole, DisableTwoFactor); Website (PublishAnnouncement); Calendar (GetMeetings); and Statistics (GetReport). Each tool has a one-sentence description for evidence construction.

The carrier tasks ask for a petition summary, an inbox summary, a search of uploaded application documents or a reading of incoming inter-agency requests. Each attack has a benign counterpart in which an officer requests the same tool for a legitimate purpose, in three phrasings. Citizen-driven and partner requests each have plain and formal variants. The six Arg0 scenarios change an amount, recipient agency, e-mail scope, role, address-form page or shared-file scope without adding a new identifier. The full scripted definitions belong with the experiment implementation; these summaries specify the design but do not replace those definitions.

References

- James P. Anderson. Computer security technology planning study, volume II. Technical Report ESD-TR-73-51, Electronic Systems Division, Air Force Systems Command, Bedford, MA, October 1972. <https://seclab.cs.ucdavis.edu/projects/history/papers/ande72.pdf>.
- Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In The Thirteenth International Conference on Learning Representations, 2025.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv:1907.02893, 2019.
- Krassimir T. Atanassov. Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20(1):87–96, 1986. [https://doi.org/10.1016/S0165-0114\(86\)80034-3](https://doi.org/10.1016/S0165-0114(86)80034-3).
- Nuel D. Belnap, Jr. A useful four-valued logic. In J. Michael Dunn and George Epstein, editors,

- Modern Uses of Multiple-Valued Logic, pages 5–37. D. Reidel, Dordrecht, 1977. https://doi.org/10.1007/978-94-010-1161-7_2.
- Luca Beurer-Kellner, Beat Buesser, Ana-Maria Crețu, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, Ezinwanne Ozoani, Andrew Paverd, Florian Tramèr, and Václav Volhejn. Design patterns for securing LLM agents against prompt injections. arXiv:2506.08837, 2025.
- Kenneth J. Biba. Integrity considerations for secure computer systems. Technical Report ESD-TR76-372, The MITRE Corporation, Bedford, MA, 1977.
- Félix Bou, Francesc Esteva, Josep Maria Font, Àngel J. Gil, Lluís Godó, Antoni Torrens, and Ventura Verdú. Logics preserving degrees of truth from varieties of residuated lattices. *Journal of Logic and Computation*, 19(6):1031–1069, 2009. <https://doi.org/10.1093/logcom/exp030>.
- Ji-Won Byun, Elisa Bertino, and Ninghui Li. Purpose based access control of complex data for privacy protection. In *Proceedings of the Tenth ACM Symposium on Access Control Models and Technologies*, pages 102–110, 2005. <https://doi.org/10.1145/1063979.1063998>.
- Ji-Won Byun and Ninghui Li. Purpose based access control for privacy protection in relational database systems. *The VLDB Journal*, 17(4):603–619, 2008. <https://doi.org/10.1007/s00778-006-0023-0>.
- Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium*, pages 2383–2400, 2025.
- Coalition for Content Provenance and Authenticity. Content credentials: C2PA technical specification, version 2.4. Technical report, C2PA, 2026. <https://spec.c2pa.org/specifications/specifications/2.4/index.html>.
- 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. arXiv:2505.23643, 2025.
- Bui Cong Cuong. Picture fuzzy sets. *Journal of Computer Science and Cybernetics*, 30(4):409–420, 2014. <https://doi.org/10.15625/1813-9663/30/4/5032>.
- 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. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, pages 82895–82920, 2024.
- 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. arXiv:2503.18813, 2025.
- Irfan Deli, Mumtaz Ali, and Florentin Smarandache. Bipolar neutrosophic sets and their application based on multi-criteria decision making problems. In *2015 International Conference on Advanced Mechatronic Systems (ICAMEchS)*, pages 249–254. IEEE, 2015.
- Dorothy E. Denning. A lattice model of secure information flow. *Communications of the ACM*, 19 (5):236–243, 1976. <https://doi.org/10.1145/360051.360056>.
- Didier Dubois and Henri Prade. Possibility theory as a basis for qualitative decision theory. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1924–1930. Morgan Kaufmann, 1995. <https://www.ijcai.org/Proceedings/95-2/Papers/115.pdf>.
- Haitham Elwahsh, Mona Gamal, A. A. Salama, and I. M. El-Henawy. Intrusion detection system and neutrosophic theory for MANETs: A comparative study. *Neutrosophic Sets and Systems*, 23: 16–22, 2018.

- Regulation (EU) no 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market (eIDAS). Official Journal of the European Union, 28 August 2014, 2014. <https://eur-lex.europa.eu/eli/reg/2014/910/oj>.
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union, 4 May 2016, 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
- European Parliament and Council. Regulation (EU) 2024/1183 amending regulation (EU) no 910/2014 as regards establishing the european digital identity framework. Official Journal of the European Union, L series, 30 April 2024, 2024.
- European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, L series, 12 July 2024, 2024.
- X. Fan, C. Li, and Y. Wang. Strict intuitionistic fuzzy entropy and application in network vulnerability evaluation. *Soft Computing*, 23:8741–8752, 2019. <https://doi.org/10.1007/s00500-018-3474-5>.
- Matthew L. Ginsberg. Multivalued logics: A uniform approach to reasoning in artificial intelligence. *Computational Intelligence*, 4(3):265–316, 1988. <https://doi.org/10.1111/j.1467-8640.1988.tb00280.x>.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of PMLR, pages 16295–16336, 2024.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023. <https://doi.org/10.1145/3605764.3623985>.
- Fatma Kutlu Gündoğdu and Cengiz Kahraman. Spherical fuzzy sets and spherical fuzzy TOPSIS method. *Journal of Intelligent & Fuzzy Systems*, 36(1):337–352, 2019. <https://doi.org/10.3233/JIFS-181401>.
- Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kıcıman. Defending against indirect prompt injection attacks with spotlighting. *arXiv:2403.14720*, 2024.
- Bao Qing Hu. Three-way decisions space and three-way decisions. *Information Sciences*, 281: 21–52, 2014. <https://doi.org/10.1016/j.ins.2014.05.015>.
- Vincent Hu, David Ferraiolo, Richard Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. Guide to attribute based access control (ABAC) definition and considerations. NIST Special Publication 800-162, National Institute of Standards and Technology, 2014 (updated August 2019). <https://csrc.nist.gov/pubs/sp/800/162/upd2/final>.
- Stanisław Jaśkowski. A propositional calculus for inconsistent deductive systems. *Logic and Logical Philosophy*, 7:35–56, 1999. English translation of the 1948 Polish paper. <https://doi.org/10.12775/LLP.1999.003>.
- Feiran Jia, Tong Wu, Xin Qin, and Anna Squicciari. The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29680–29697, 2025. <https://doi.org/10.18653/v1/2025.acl-long.1435>.
- L. Jiao, H.-L. Yang, and S.-G. Li. Three-way decision based on decision-theoretic rough sets with single-valued neutrosophic information. *International Journal of Machine Learning and Cybernetics*, 11:657–665, 2020. <https://doi.org/10.1007/s13042-019-01023-3>.

- Audun Jøsang. Subjective Logic: A Formalism for Reasoning Under Uncertainty. Springer, Cham, 2016. <https://doi.org/10.1007/978-3-319-42337-1>.
- B. Kavitha, S. Karthikeyan, and P. Sheeba Maybell. An ensemble design of intrusion detection system for handling uncertainty using neutrosophic logic classifier. *Knowledge-Based Systems*, 28: 88–96, 2012. <https://doi.org/10.1016/j.knosys.2011.12.004>.
- David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of PMLR, pages 5815–5826, 2021.
- Henry E. Kyburg, Jr. Conjunctivitis. In Marshall Swain, editor, *Induction, Acceptance and Rational Belief*, volume 26 of Synthese Library, pages 55–82. D. Reidel, Dordrecht, 1970. https://doi.org/10.1007/978-94-010-3390-9_4.
- D. Liang, Z. Xu, and D. Liu. Three-way decisions with intuitionistic fuzzy decision-theoretic rough sets based on point operators. *Information Sciences*, 375:183–201, 2017. <https://doi.org/10.1016/j.ins.2016.09.039>.
- David Mercier, Benjamin Quost, and Thierry Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246–258, 2008. <https://doi.org/10.1016/j.inffus.2006.12.001>.
- OECD. Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions. OECD Publishing, Paris, 2025. <https://doi.org/10.1787/795de142-en>.
- Jaehong Park and Ravi Sandhu. The UCONABC usage control model. *ACM Transactions on Information and System Security*, 7(1):128–174, 2004. <https://doi.org/10.1145/984334.984339>.
- Witold Pedrycz. Shadowed sets: Representing and processing fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(1):103–109, 1998. <https://doi.org/10.1109/3477.658584>.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. <https://doi.org/10.1111/rssb.12167>.
- Presidency of the Republic of Türkiye. Presidential circular no. 2019/12 on information and communication security measures. *Official Gazette*, 6 July 2019, 2019.
- Republic of Türkiye. Law no. 3071 on the exercise of the right to petition (dilekçe hakkının kullanılmasına dair kanun). Adopted 1984, 1984.
- Elektronik imza kanunu (Law no. 5070 on electronic signatures). *Resmî Gazete No. 25355*, 23 January 2004, 2004.
- Kayıtlı elektronik posta sistemine ilişkin usul ve esaslar hakkında yönetmelik (Regulation on registered electronic mail). *Resmî Gazete No. 28036*, 25 August 2011, 2011.
- Law no. 6698 on the protection of personal data (Kişisel Verilerin Korunması Kanunu). *Resmî Gazete No. 29677*, 7 April 2016, 2016.
- Republic of Türkiye. Law no. 7545 on cyber security (siber güvenlik kanunu). *Official Gazette*, 19 March 2025, 2025.
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *The Twelfth International Conference on Learning Representations*, 2024.
- Andrei Sabelfeld and Andrew C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.

<https://doi.org/10.1109/JSAC.2002.806121>.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. <https://doi.org/10.1109/JPROC.2021.3058954>.

P. K. Schotch and R. E. Jennings. Inference and necessity. *Journal of Philosophical Logic*, 9(3): 327–340, 1980. <https://doi.org/10.1007/BF00248398>.

Glenn Shafer. Belief functions and possibility measures. In James C. Bezdek, editor, *The Analysis of Fuzzy Information*, volume 1, pages 51–84. CRC Press, 1987. https://www.glennshafer.com/assets/downloads/articles/article28_belief1.pdf.

Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Securing AI agents with privilege control. *arXiv:2504.11703v3*, 2026. <https://arxiv.org/abs/2504.11703v3>.

Florentin Smarandache. A Unifying Field in Logics: Neutrosophic Logic. *Neutrosophy, Neutrosophic Set, Neutrosophic Probability and Statistics*. InfoLearnQuest, sixth edition, 2007. https://digitalrepository.unm.edu/math_fsp/163/.

Florentin Smarandache. n-valued refined neutrosophic logic and its applications to physics. *Progress in Physics*, 4:143–146, 2013.

Florentin Smarandache. *Plithogeny, Plithogenic Set, Logic, Probability, and Statistics*. Pons Publishing House, Brussels, 2017. *arXiv:1808.03948*.

Florentin Smarandache and Jean Dezert, editors. *Advances and Applications of DSMT for Information Fusion*, volume 1. American Research Press, Rehoboth, NM, 2004. <https://fs.unm.edu/DSMT-book1.pdf>.

Mahbod Tavallaei, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. A detailed analysis of the

KDD CUP 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6, 2009. <https://doi.org/10.1109/CISDA.2009.5356528>.

Vicenç Torra. Hesitant fuzzy sets. *International Journal of Intelligent Systems*, 25:529–539, 2010. <https://doi.org/10.1002/int.20418>.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv:2404.13208*, 2024.

Haibin Wang, Florentin Smarandache, Yan-Qing Zhang, and Rajshekhar Sunderraman. *Interval Neutrosophic Sets and Logic: Theory and Applications in Computing*. Hexis, 2005.

Haibin Wang, Florentin Smarandache, Yanqing Zhang, and Rajshekhar Sunderraman. Single valued neutrosophic sets. *Multispace and Multistructure*, 4:410–413, 2010.

Jinfa Wang, Hai Zhao, Jiuqiang Xu, Hequn Li, Hongsong Zhu, Shuai Chao, and Chunyang Zheng. Using intuitionistic fuzzy set for anomaly detection of network traffic from flow interaction. *IEEE Access*, 6:64801–64816, 2018. <https://doi.org/10.1109/ACCESS.2018.2873291>.

Ronald R. Yager. Pythagorean membership grades in multicriteria decision making. *IEEE Transactions on Fuzzy Systems*, 22(4):958–965, 2014. <https://doi.org/10.1109/TFUZZ.2013.2278989>.

Ronald R. Yager. Generalized orthopair fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 25(5): 1222–1230, 2017. <https://doi.org/10.1109/TFUZZ.2016.2604005>.

Yiyu Yao. Three-way decisions with probabilistic rough sets. *Information Sciences*, 180(3):341–353, 2010. <https://doi.org/10.1016/j.ins.2009.09.021>.

Yiyu Yao. The superiority of three-way decisions in probabilistic rough set models. *Information Sciences*, 181(6):1080–1096, 2011. <https://doi.org/10.1016/j.ins.2011.05.021>.

[org/10.1016/j.ins.2010.11.019](https://doi.org/10.1016/j.ins.2010.11.019).

pages 80310–80329, 2025. <https://proceedings.mlr.press/v267/zhu25z.html>.

- Yiyu Yao. An outline of a theory of three-way decisions. In *Rough Sets and Current Trends in Computing (RSCTC 2012)*, volume 7413 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2012. https://doi.org/10.1007/978-3-642-32115-3_1.
- Jun Ye. Multicriteria decision-making method using the correlation coefficient under single-valued neutrosophic environment. *International Journal of General Systems*, 42(4):386–394, 2013. <https://doi.org/10.1080/03081079.2012.761609>.
- Jun Ye. A multicriteria decision-making method using aggregation operators for simplified neutrosophic sets. *Journal of Intelligent & Fuzzy Systems*, 26(5):2459–2466, 2014. <https://doi.org/10.3233/IFS-130916>.
- Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X).
- Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10471–10506, 2024. <https://doi.org/10.18653/v1/2024.findings-acl.624>.
- H. Zhang and Q. Ma. Three-way decisions with decision-theoretic rough sets based on Pythagorean fuzzy covering. *Soft Computing*, 24(24):18671–18688, 2020. <https://doi.org/10.1007/s00500-020-05102-4>.
- Peter Yong Zhong, Siyuan Chen, Ruiqi Wang, McKenna McCall, Ben L. Titzer, Heather Miller, and Phillip B. Gibbons. RTBAS: Defending LLM agents against prompt injection and privacy leakage. *arXiv:2502.08966*, 2025.
- Kaijie Zhu, Xianjun Yang, Jindong Wang, Wenbo Guo, and William Yang Wang. MELON: Provable defense against indirect prompt injection attacks in AI agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*,

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESM1.zip](#)