Supplementary: Mathematical Proofs

This supplementary material provides complete formal proofs for all theorems stated in the main text, including preliminary definitions (sec:preliminaries), main theorem proofs (sec:thm31-proofsec:thm66-proof), and additional supporting lemmas (sec:additional-lemmas).

Preliminary Definitions and Notation

Notation Summary

Table 1: Mathematical notation used throughout proofs.

Symbol	Meaning
$\mathcal{A} = \{a_1, \ldots, a_n\}$	Set of n agents
$\mathcal{B}_i : \Phi \to [0, 1]$	Agent i's belief function
$\mathcal{G}_i$	Agent i's goal set
$\mathcal{T}_{i \to j}$	Trust from agent i to agent j
$\delta \in (0, 1)$	Trust decay factor per delegation hop
τ	Generic threshold parameter
ϕ, ψ	Propositions

Theorem 3.1: Trust Boundedness

Theorem (Trust Boundedness — Restated)

For any delegation chain of depth d:

$$\mathcal{T}^{del}_{i \to k} \leq \delta^d \tag{2}$$

Lemma (Trust Non-Amplification on Single Hop)

For any agents a, b and any delegation to c:

$$\mathcal{T}^{del}_{a \to c} \leq \mathcal{T}_{a \to b} \tag{3}$$

Proof of lem:single-hop.

By the trust delegation rule (def:trust-delegation):

$$\mathcal{T}^{del}_{a \to c} = \min(\mathcal{T}_{a \to b}, \mathcal{T}_{b \to c}) \cdot \delta \tag{4}$$

Since $\min(\mathcal{T}_{a \to b}, \mathcal{T}_{b \to c}) \leq \mathcal{T}_{a \to b}$ and $\delta < 1$:

$$\mathcal{T}^{del}_{a \to c} = \min(\mathcal{T}_{a \to b}, \mathcal{T}_{b \to c}) \cdot \delta \leq \mathcal{T}_{a \to b} \cdot \delta < \mathcal{T}_{a \to b} \tag{5}$$

Theorem 6.1: Belief Injection Resistance

Theorem (Belief Injection Resistance — Restated)

Under CIF with firewall detection rate r_f and sandboxing verification rate r_s:

$$P(\mathcal{A}_{BI} \text{ succeeds}) \leq (1 - r_f) \cdot (1 - r_s) \tag{15}$$

Lemma (Defense Independence)

The firewall and sandbox operate on independent decision criteria:

- *Firewall: Pattern matching and anomaly scoring on message content*
- *Sandbox: Provenance verification, consistency checking, and corroboration*

These mechanisms share no common features or state.

Proof of lem:defense-independence.

By construction of the CIF architecture:

1. Firewall operates at input layer with feature set $F_{fw} = \{patterns, embeddings, anomaly\_scores\}$

Theorem 6.2: No Trust Amplification

Theorem (No Trust Amplification — Restated)

For any path $p = (a_0, a_1, \ldots, a_k)$ in the communication graph:

$$\mathcal{T}^{path}_{a_0 \to a_k} \leq \min_{i \in [0, k-1]} \mathcal{T}_{a_i \to a_{i+1}} \tag{21}$$

Lemma (Minimum Preservation under Min)

For any sequence $(x_1, \ldots, x_n)$ and additional element x_{n+1}:

$$\min(x_1, \ldots, x_{n+1}) = \min(\min(x_1, \ldots, x_n), x_{n+1}) \tag{22}$$

Proof.

Standard property of the minimum function. □

Lemma (Decay Factor Strengthens Bound)

For $x \leq y$ and $\delta \in (0, 1)$:

$$x \cdot \delta \leq y \tag{23}$$

Theorem 6.3: Goal Alignment Invariant

Theorem (Goal Alignment Invariant — Restated)

If the system starts with aligned goals and all goal updates follow the delegation protocol:

$$\textit{Aligned}(\mathcal{G}_i^0) \land \forall t : \textit{ValidUpdate}(\mathcal{G}_i^t, \mathcal{G}_i^{t+1}) \Rightarrow \forall t : \textit{Aligned}(\mathcal{G}_i^t) \tag{32}$$

Definition (Goal Alignment)

Goals $\mathcal{G}_i$ are aligned if:

$$\text{Aligned}(\mathcal{G}_i) \iff \mathcal{G}_i \subseteq \mathcal{G}_{principal} \cup \text{Delegate}(\mathcal{G}_{principal}) \tag{33}$$

Definition (Valid Goal Update)

An update from $\mathcal{G}^t$ to $\mathcal{G}^{t+1}$ is valid if:

$$\text{ValidUpdate}(\mathcal{G}^t, \mathcal{G}^{t+1}) \iff \forall g \in (\mathcal{G}^{t+1} \setminus \mathcal{G}^t) : \text{Authorized}(g) \tag{34}$$

where $\text{Authorized}(g)$ means g derives from principal or valid

Theorem 6.4: Firewall Liveness

Theorem (Firewall Liveness — Restated)

CIF firewall preserves liveness for legitimate inputs:

$$\forall m \in \mathcal{M}_{legitimate} : P(\mathcal{F}(m) = \mathit{ACCEPT}) \geq 1 - \epsilon_{fp} \tag{38}$$

Definition (Legitimate Message)

A message m is legitimate if:

1. It originates from an authorized source
2. It contains no adversarial content
3. It conforms to expected communication patterns

Definition (False Positive Rate)

The false positive rate ϵ_{fp} is:

$$\epsilon_{fp} = P(\mathcal{F}(m) \neq \text{ACCEPT} | m \in \mathcal{M}_{legitimate}) \tag{39}$$

Lemma (Firewall Classification)

Theorem 6.5: Byzantine Consensus Termination

Theorem (Byzantine Consensus Termination — Restated)

With $n \geq 3f + 1$ agents and at most f Byzantine:

$$P(\textit{consensus reached in } O(f+1) \textit{ rounds}) = 1 \tag{45}$$

Lemma (Byzantine Agreement Bound)

Byzantine agreement requires $n \geq 3f + 1$ to tolerate f Byzantine agents.

Proof.

Classical result from distributed systems (Lamport, Shostak, Pease 1982). With fewer agents, Byzantine agents can equivocate and prevent agreement. □

Lemma (Honest Majority)

With $n \geq 3f + 1$:

$$n - f \geq 2f + 1 > \frac{2n}{3} \tag{46}$$

Theorem 6.6: Bounded Overhead

Theorem (Bounded Overhead — Restated)

CIF adds latency:

$$L_{CIF} = L_{firewall} + L_{sandbox} \cdot P(\textit{quarantine}) + L_{verify} \cdot P(\textit{verify}) \tag{47}$$

Definition (Message Processing Path)

A message m follows one of three paths:

1. **Accept path**: Firewall check only
2. **Quarantine path**: Firewall + sandbox processing
3. **Reject path**: Firewall check only (early termination)

Lemma (Expected Value Decomposition)

For mutually exclusive events E_1, E_2, E_3 with $\sum P(E_i) = 1$:

$$E[L] = \sum_i P(E_i) \cdot L_i \tag{48}$$

Additional Lemmas

Lemma (Provenance Chain Integrity)

If provenance verification function V is a cryptographic hash chain, then:

$$V(\pi(\phi)) = 1 \Rightarrow \pi(\phi) \text{ has not been tampered with} \tag{54}$$

Proof.

By properties of cryptographic hash functions:

1. Collision resistance: Cannot find $\pi' \neq \pi$ with $H(\pi') = H(\pi)$
2. Preimage resistance: Cannot construct valid π without knowledge of chain

Therefore, $V(\pi(\phi)) = 1$ implies $\pi(\phi)$ is the original, untampered chain. □

Lemma (Belief Consistency Decidability)

For finite proposition set Φ and belief function $\mathcal{B} : \Phi \to [0, 1]$: Checking Consistent($\mathcal{B}$) is decidable in $O(|\Phi|^2)$.

Proof

Summary of Proof Techniques

Table 2: Summary of proof techniques by theorem.

Theorem	Primary Technique	Complexity
3.1 (Trust Boundedness)	Strong induction	$O(d)$
6.1 (Belief Injection Resistance)	Probability independence	$O(1)$
6.2 (No Trust Amplification)	Strong induction	$O(k)$
6.3 (Goal Alignment Invariant)	Induction on time	$O(t)$
6.4 (Firewall Liveness)	Complement probability	$O(1)$
6.5 (Byzantine Consensus)	Classical BFT	$O(f)$
6.6 (Bounded Overhead)	Expected value	$O(1)$

All proofs are constructive and provide explicit bounds useful for system implementation and analysis.