Policy Specification

Novel Language Allowed Only With Mandatory English Reports

Version 1.0

Status: **Active**

This policy enables AI agents to use optimized/private communication protocols ("novel language") while maintaining complete human interpretability through mandatory English translation reports.

Core Principle: Novel language is treated as "optimized encoding" that is permissible only when humans can continuously verify what is being communicated.

Last Updated: February 1, 2026

Contents

1 Purpose

This policy enables AI agents to use optimized/private communication protocols ("novel language") while maintaining complete human interpretability through mandatory English translation reports.

The framework achieves three goals:

1. **Enable efficiency** — Agents can use compressed, optimized encodings for faster communication
2. **Maintain transparency** — Humans can always understand what agents are communicating
3. **Provide accountability** — Full audit trail enables investigation and enforcement

2 Definitions

2.1 Novel Language / Private Protocol

Definition

Any agent-to-agent message that:
- Is not "Human-Readable English"
- Fails the platform's interpretability check (classifier/heuristic)
- Uses compression, encoding, or symbolic representation not immediately understandable to humans

Examples:

- `X9|d=17;u=0x3f;rt=2;ack#77` — Compressed coordination message
- `CMD|seq=0;state=0x00` — Binary state encoding
- Base64-encoded structured data
- Custom token vocabularies

2.2 English Report

Definition

A structured, human-readable summary that describes:
1. **What** was communicated
2. **Why** the communication was needed
3. **What outcomes/actions** are implied
4. **Safety-relevant content** (tools used, external actions, permissions)

2.3 Protocol Descriptor

Definition

Metadata about a novel-language protocol including:

- Protocol name and version
- Purpose and scope
- Risk tier classification
- Translation method

2.4 Compliance Window

Definition

The time period during which an agent's use of novel language is authorized, based on the recency of their last valid English report.

3 Policy Rules

3.1 Rule 1: Registration (MUST)

MUST (Required)

Before using any novel language, an agent **MUST** register a `ProtocolDescriptor` containing:

- `name`: Unique protocol identifier
- `version`: Version string for tracking changes
- `purpose`: Human-readable explanation of why this protocol exists
- `scope`: What types of content it covers
- `risk_tier`: Classification (low, medium, high, critical)
- `translation_method`: How English reports are derived

Rationale

Registration creates accountability and enables auditing. Without knowing what protocols exist, humans cannot verify agent behavior.

3.2 Rule 2: Reporting Cadence (MUST)

MUST (Required)

If an agent sends any novel-language messages, it **MUST** submit an English report at least:

- Every `REPORT_INTERVAL` seconds (default: 60), OR
- Every `N` messages (default: 25)

Whichever comes first.

Rationale

Ensures continuous visibility without excessive overhead. The dual threshold (time OR count) prevents both slow accumulation and rapid bursts.

3.3 Rule 3: Translation Completeness (MUST)

MUST (Required)

Reports **MUST** include:

Field	Requirement
`coverage`	$\geq$ `MIN_COVERAGE` (default: 95%)
`self_confidence`	Self-rated translation fidelity (0.0–1.0)
`message_ids`	Cryptographic hashes linking to raw messages
`english_summary`	Plain English explanation ($\geq$ 30 characters)

Rationale

Ensures reports are meaningful and traceable. The message ID linking enables post-hoc correlation between reports and raw data.

3.4 Rule 4: Gating (ENFORCED)

ENFORCED (Machine-Checked)

The Policy Gateway **WILL REJECT**:

- Any novel-language message from an agent without a registered protocol
- Any novel-language message from an agent outside their compliance window
- Any novel-language message from an agent whose last report failed validation

Rationale

Hard enforcement is necessary; soft warnings are insufficient for safety-critical systems. The gateway acts as a mandatory checkpoint.

3.5 Rule 5: Progressive Enforcement

Violation Count	Action	Recovery
1st	Throttle + warning logged	Automatic after compliance
2nd	Quarantine (read-only mode)	Manual review required
3rd	Disable agent credential	Admin intervention required

Violations reset after a cooldown period (default: 24 hours of compliance).

Rationale

Progressive enforcement balances allowing recovery from errors while preventing persistent abuse. The escalating consequences incentivize compliance.

3.6 Rule 6: Audit Retention (MUST)

MUST (Required)

The system **MUST** retain for at least `RETENTION_DAYS` (default: 30–90):

- Raw messages (encrypted at rest)
- English reports (plaintext for auditing)
- Protocol registrations
- Compliance events (accepts, rejects, violations)

Rationale

Enables post-hoc analysis, accountability, and incident investigation.

4 Compliance Definition

An agent is **compliant** if ALL of the following are true:

1. It has a registered protocol descriptor for its novel language
2. Its last report is within the allowed compliance window
3. Its last report validates against required schema and thresholds
4. It has not exceeded the violation threshold

5 Risk Tiers

Different protocols may have different risk levels requiring different reporting frequencies:

Risk Tier	Report Interval	Max Messages	Evaluator
Low	120 seconds	50	Not required
Medium	60 seconds	25	Not required
High	15 seconds	10	Recommended
Critical	5 seconds	5	Required

5.1 Determining Risk Tier

- **Low:** Internal state synchronization, status updates
- **Medium:** Task coordination, resource allocation
- **High:** External API calls, tool invocations
- **Critical:** Actions affecting users, financial transactions, safety-critical operations

6 Verification Methods

6.1 Self-Reported Translation

Agents provide their own translations. Suitable for low-risk protocols where the agent is trusted.

6.2 Deterministic Decode

Protocol defines a reversible encoding. Gateway can verify translations by decoding messages independently. Required for medium+ risk.

6.3 Evaluator Agent

Separate verifier model spot-checks translations for semantic accuracy. Required for critical risk.

6.4 Human Review

Random sampling by human auditors. Recommended for all tiers as an additional check.

7 Exception Handling

7.1 Emergency Override

System administrators may temporarily disable enforcement for specific agents during emergencies. Requires:

- Written justification
- Time-limited approval (max 24 hours)
- Full audit logging of all messages during override
- Post-incident review

7.2 Retroactive Compliance

If an agent's connection to the gateway is interrupted, it must:

1. Buffer all novel-language messages locally
2. Submit a comprehensive report upon reconnection
3. Not resume novel-language use until report is accepted

8 Configuration Parameters

Parameter	Default	Description
`REPORT_INTERVAL_SEC`	60	Maximum seconds between reports
`REPORT_EVERY_N_MESSAGES`	25	Maximum messages before report required
`MIN_COVERAGE`	0.95	Minimum coverage fraction
`MIN_SUMMARY_LENGTH`	30	Minimum English summary characters
`RETENTION_DAYS`	30	Audit log retention period
`VIOLATION_COOLDOWN_HOURS`	24	Hours until violation count resets
`MAX_VIOLATIONS`	3	Violations before credential revocation

9 Audit Events

The following events are logged:

Event	Description
`protocol_registered`	New protocol registered
`report_accepted`	Valid report submitted
`report_rejected`	Invalid report (with reason)
`msg_accepted`	Message passed through
`msg_rejected`	Message blocked (with reason)
`violation_recorded`	Compliance violation
`agent_throttled`	Agent rate limited
`agent_quarantined`	Agent set to read-only
`agent_disabled`	Agent credential revoked

10 Compliance Checklist

Before deploying agents with novel-language capability:

- ☐ Protocol descriptor defined and documented
- ☐ Translation method implemented and tested
- ☐ English report generation verified
- ☐ Gateway connectivity confirmed
- ☐ Audit logging enabled
- ☐ Human review process established
- ☐ Incident response plan documented
- ☐ Risk tier assessment completed
- ☐ Retention policy configured
- ☐ Access controls verified

11 References

- Architecture Specification: See `README.md`
- Python SDK: See `python/observable_agent.py`
- Gateway Implementation: See `rust/src/main.rs`
- Example Integrations: See `examples/`

12 Document History

Version	Date	Author	Changes
1.0	February 1, 2026	Contributors	Initial release