# DUALNET: A Cryptographically-Mediated Separation Protocol for Human and AI Internets

Constantinescu Mihnea Cristian

**Abstract**

A complete separation between a network of autonomous AI agents and a network of human users would let humans benefit from AI computation without exposing the human internet to agent-initiated interaction. Existing isolation mechanisms — anonymity networks, air gaps, unidirectional data diodes, and trusted execution environments — each address a different problem: they hide identities, sever connectivity entirely, restrict physical flow direction, or protect computation, but none provides interactive yet asymmetric communication between two populations of actors. We propose DUALNET, a layered architecture in which two parallel networks operate under incompatible protocols and are connected by a single mediating gateway. Information flow from humans to agents requires a zero-knowledge proof of personhood and a time-bounded cryptographic mandate; flow from agents to humans requires a valid return token bound to such a mandate. We specify L-AI, a constructed agent-side language over a symbol space disjoint from human scripts, whose grammar lacks the type required to reference human individuals, and whose ingress filter rejects inputs statistically consistent with human language. We give a formal $(q, \varepsilon)$-separation definition, analyze the construction under correlated-layer adversaries, bound covert-channel bandwidth, and discuss the governance problem that arises when the temporary AI authority that designed L-AI is terminated. We state our assumptions and limitations explicitly: the linguistic filter inherits the adversarial-robustness limits of statistical classifiers, and long-term security reduces to the integrity of the institution that inherits the system.

## 1. Introduction

The contemporary internet is a shared medium for human and machine actors. Distinguishing the two has become a recurring problem: web services deploy bot detection, CAPTCHA challenges [6], and behavioral fingerprinting to identify automated traffic [1], while autonomous agents increasingly operate as legitimate users on services designed for humans [2]. The asymmetry of cost between producing a request and verifying its origin creates a structural vulnerability: an attacker who controls a population of agents can flood, deceive, or extract from systems whose access controls assume human-scale interaction rates. This is the network analogue of the Sybil attack [7]: without a binding between identities and scarce real-world resources, one adversary can present as many.

What is needed is a system in which the human internet and the agent internet are not merely distinguishable, but architecturally separate — while preserving the ability of humans to query agents and receive answers. The required property is therefore not isolation but *asymmetric interactivity*: human-initiated round trips succeed; agent-initiated contact fails.

We propose DUALNET, a protocol stack that achieves this separation through five composable mechanisms: (i) proof-of-personhood at the human gateway, (ii) incompatible transport protocols on the two sides, (iii) a single mediated translation gateway, (iv) capability tokens that lose validity outside the AI network, and (v) behavioral anomaly detection as a final tripwire. We further specify a constructed AI-only language, L-AI, designed by a temporary supervised AI authority and frozen at

deployment, and we show how it raises the cost of cross-network attacks under explicitly stated assumptions.

Our contributions are: (1) a formal definition of asymmetric network separation (Section 3); (2) the DUALNET architecture and the mandated translation protocol (Sections 4–7); (3) a security analysis that does not assume layer independence and instead derives upper and lower breach-probability bounds (Section 8); (4) a construction sketch for L-AI with an explicit account of its adversarial limitations (Section 6); and (5) an analysis of post-deployment governance, including three succession models for the terminated design authority (Section 9). Section 10 states limitations and open problems.

## 2. Related Work

DUALNET sits at the intersection of network isolation, identity verification, and information-flow control. We position it against five families of prior work.

**Anonymity networks.** Mix networks [8] and onion routing systems such as Tor [9] address *who is communicating*, not *what kind of actor* is communicating. They deliberately make human and automated traffic indistinguishable — the opposite of our goal. DUALNET requires actor-class attribution at the gateway while preserving individual anonymity within the human class; we achieve this combination through zero-knowledge personhood proofs rather than routing indirection.

**Air-gapped systems.** Physical isolation severs connectivity entirely and is the standard for high-assurance military and industrial networks. Air gaps provide no interactivity: any sanctioned crossing (removable media, manual transfer) reintroduces the full attack surface, as demonstrated repeatedly by malware that traverses air gaps via USB devices. DUALNET can be read as an attempt to formalize what a *safe, interactive* crossing would require.

**Unidirectional gateways (data diodes).** Hardware data diodes enforce one-way information flow at the physical layer and are deployed in critical infrastructure. They are the closest prior art to our asymmetry goal, but they operate below semantics: a diode permits all traffic in one direction and none in the other, whereas DUALNET must permit *responses* to flow agent-to-human while blocking agent-*initiated* messages in the same direction. This requires session-level state (the mandate and return token of Section 7), which no physical-layer mechanism can express.

**Trusted execution environments.** Enclaves such as Intel SGX [10] protect the integrity and confidentiality of *computation* against a hostile host. DUALNET addresses the dual problem: protecting a *population of hosts* (humans) against hostile computation. We borrow the TEE notion of attestation for AINet agents (Section 4) but apply it to network admission rather than code identity.

**Proof of personhood and bot detection.** CAPTCHAs [6] impose a per-request human-effort cost but have been progressively defeated by machine perception. Personhood-credential proposals [4, 11] bind network identities to scarce real-world enrollment events, directly addressing the Sybil problem [7]. DUALNET adopts this line of work as its first layer and contributes the surrounding architecture: personhood credentials answer "is this a human?" but not "how do we let humans query agents without letting agents address humans?" — the question the remaining four layers answer.

Finally, the covert-channel limits we inherit are classical: Lampson's confinement problem [12] established that a determined insider can leak information through resource-usage patterns, and our analysis in Section 8.3 treats this as a bandwidth-bounding problem rather than an eliminable one.

## 3. Threat Model and Security Definition

We consider an adversary *A* that controls a population of autonomous agents operating within AINet. *A*'s objective is to deliver a message m to a human user u on HumanNet without u having initiated a corresponding query. We assume:

1. *A* has large but bounded computational resources within AINet and cannot break standard cryptographic primitives (EUF-CMA signatures, collision-resistant hashes, pseudorandom MACs).

2. *A* cannot forge personhood credentials issued by honest attestation authorities; at most a fraction f < 1 of authorities are corrupted, and no single authority suffices to issue a valid credential.

3. The Translation Gate G is honest-but-auditable: G follows its specification, and a compromise of G is a system-wide failure handled by governance (Section 9), not by the protocol.

4. Covert channels in timing, volume, and frequency exist and cannot be eliminated [12]; they can only be rate-bounded.

We can now state the goal formally.

> **Definition 1 ((q, ε)-separation).** A two-network system with gateway G provides *(q, ε)-separation* if every adversary *A* satisfying assumptions 1–4, making at most q delivery attempts, causes an unsolicited agent-originated message to be rendered to a human user with probability at most ε, where the probability is over the randomness of the protocol and of *A*.

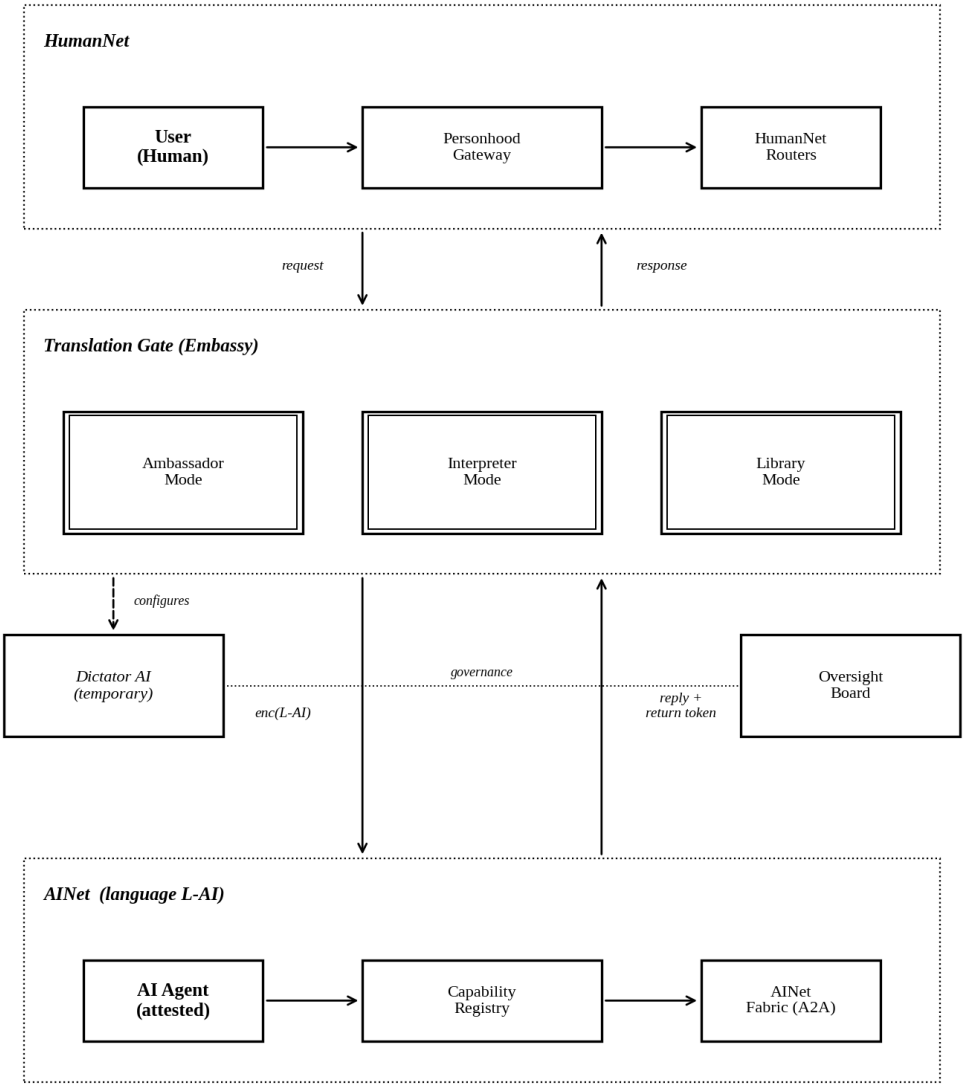
Two remarks. First, the definition deliberately excludes covert channels to human *collaborators*: if a human chooses to decode timing patterns, that human has initiated contact, which is outside the property we protect. Covert channels are instead treated quantitatively in Section 8.3. Second, ε is per-attempt-budget, not per-attempt: the definition composes over an attack campaign, which is the operationally meaningful unit.

## 4. Architecture Overview

DUALNET consists of three zones connected linearly. HumanNet contains human users and the routers and services they interact with, operating under a modified HTTP variant denoted HTTP-H. AINet contains attested AI agents, a Capability Registry, and an inter-agent transport fabric operating under an Agent-to-Agent protocol denoted A2A. Between them sits the Translation Gate, the sole permitted point of inter-zone communication. Agents are admitted to AINet through attestation of their software identity, analogous to enclave attestation [10], and receive capability tokens from the Registry that are valid only within AINet.

The architecture is asymmetric by design. Figure 1 illustrates the three zones, their internal components, and the directional flows between them. The Translation Gate is configured at system initialization by a temporary AI authority (the Dictator AI) and is thereafter governed by an Oversight Board modeled on international scientific consortia such as CERN [3].

*Figure 1: DUALNET Architecture*




*Asymmetric translation: human-to-AI permitted with mandate; AI-to-human requires valid return token.*

## 5. Personhood Gateway

Every connection to HumanNet must present a personhood proof. We adopt the framework of personhood credentials [4]: a user u enrolls once with a coalition of attestation authorities through a process requiring evidence not reproducible by an AI agent at scale — for example, in-person biometric registration combined with hardware-attested key generation. Requiring a threshold of authorities rather than any single one bounds the damage of insider compromise (assumption 2).

Enrollment produces a long-term secret $sk_u$ and a public credential $c_u$ proving that u is a unique human. To preserve privacy, u does not transmit $c_u$. Instead, per connection, u computes a zero-knowledge proof [13]:

$$\pi = ZKP\{ \text{knows } sk_u \text{ such that } Verify(c_u, sk_u) = 1 \}$$

where π demonstrates possession of a valid $sk_u$ without revealing which one. The HumanNet router accepts a packet if and only if π verifies and the rate-limit counter associated with π's public nullifier is within bounds. Per-credential rate limiting prevents a single compromised credential from supporting a botnet: if $r_{max}$ is the maximum sustained request rate per credential and N the number of credentials an attacker controls, attacker throughput is bounded by

$$T_{attack} \leq N \cdot r_{max},$$

and N itself is bounded by the cost of physical enrollment events, which is precisely the scarce resource the Sybil-resistance literature identifies as necessary [7].

## 6. Protocol and Linguistic Asymmetry

### *6.1. Transport-layer separation*

HTTP-H differs from standard HTTP in three respects: (a) a minimum inter-request interval $\delta_{min} \geq 50$ ms enforced at the protocol layer, (b) mandatory personhood headers on every request, and (c) content-type restrictions limiting responses to human-consumption formats. A2A is a structured binary protocol with no rendering primitives and no interval floor. An agent operating directly on HumanNet must hold a valid credential, emit human-timescale traffic, and process perceptual formats — each a constant-factor cost whose composition removes the economics of scale attacks, though not the possibility of individual ones. Layer 2 is therefore economic, not absolute.

### *6.2. The L-AI language: construction sketch*

Prior drafts of this design postulated L-AI's properties; here we sketch a construction and state what it does and does not achieve.

**Symbol space.** L-AI is defined over an alphabet Ω of tokens disjoint from every Unicode script block used by human writing systems. Disjointness is checkable mechanically at the codepoint level; an AINet ingress filter rejects any frame containing codepoints outside Ω. This check is exact, not statistical.

**Grammar.** L-AI is a typed term language. Its type system contains types for quantities, relations, capabilities, and gateway-issued opaque handles, but deliberately omits a *HumanRef* type: there is no well-typed term whose denotation is a specific human identity, postal address, or human-directed communicative act. Reference to HumanNet entities is expressible only through opaque handles minted by the Translation Gate, which alone holds the mapping. Type checking is decidable and performed at the Gate and at AINet ingress.

**Statistical filter.** Disjoint codepoints prevent verbatim human text, but an agent could encode a human-language message *within* Ω (e.g., by a substitution cipher). The ingress filter therefore additionally computes the perplexity of each frame under an ensemble of human-language models composed with candidate decodings; frames whose best decoding yields perplexity below a threshold θ under any ensemble member are rejected as suspected human-language carriers. This classifier $C_H$ is the weakest element of the construction: like all statistical classifiers it is subject to adversarial evasion [14], and we make no certified-robustness claim. Its role is to raise attack cost and to feed the audit log, not to provide a hard guarantee; the hard guarantees of the system come from the mandate mechanism

of Section 7, which does not depend on $C_H$.

L-AI is constructed once by the temporary supervised authority of Section 9 and frozen. Drift is policed by the Oversight Board through periodic statistical audits of A2A traffic against the frozen specification; the type system makes grammatical drift detectable as type errors, while lexical drift (new idioms over existing types) is the open problem discussed in Section 10.

### 7. The Mandated Translation Protocol

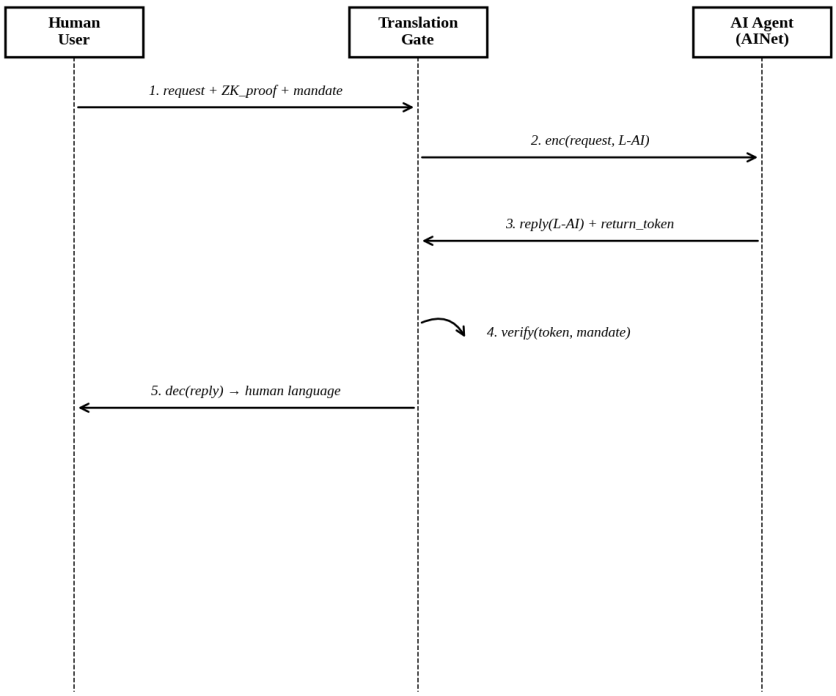
The Translation Gate G implements an asymmetric translation function. Denote by E(·) translation from a human language into L-AI and by D(·) the reverse. The asymmetry is not in the existence of these functions but in the conditions under which G applies D.

Let m be a human-language message from user u. The protocol is:

```
1. u → G  :  (m, π, mandate)
     where mandate = Sign(sk_u, H(m) || expiry || scope)

2. G verifies π and mandate. If invalid, abort.

3. G computes c = E(m) and a return_token τ:
     τ = MAC(K_G, H(mandate) || nonce)

4. G → A  :  (c, τ)  where A is the addressed agent in AINet

5. A processes c, produces reply r in L-AI

6. A → G  :  (r, τ)

7. G verifies τ against the stored mandate and checks expiry.
     If τ is valid and the mandate has not expired:
        G type-checks r, computes m' = D(r), returns m' to u.
     Otherwise:
        G discards r and logs an unauthorized return attempt.
```

The critical asymmetry is at step 7. D is well-defined and computable by G, but G refuses to apply it unless a token τ produced in step 3 of the same session is presented. Because τ is a MAC under G's key $K_G$ over H(mandate) and a fresh nonce, forging τ without compromising G contradicts MAC unforgeability (assumption 1). Since τ is issued only in response to a human-originated mandate, an agent cannot construct an originating message that G will translate: the security of the asymmetric channel reduces to standard MAC security plus the integrity of G — not to the statistical filter of Section 6.3. Figure 2 shows the message sequence.

*Figure 2: Mandated Translation Protocol*




*If step 4 fails, decryption (step 5) is refused. AI-initiated messages without a valid token never reach the human.*

## 8. Security Analysis

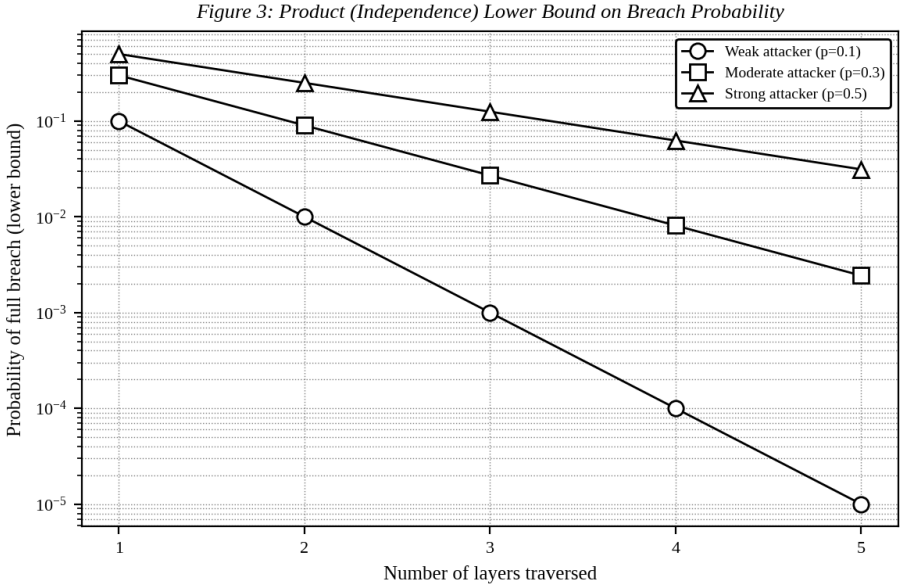
### *8.1. Breach probability without independence*

For *A* to deliver an unsolicited message to HumanNet, all five mechanisms must fail on the same attempt: personhood verification, protocol incompatibility, Gate token verification, capability validity, and behavioral detection. Let $X_i$ be the event that layer i fails, with $p_i = \Pr[X_i]$. An earlier draft assumed independence and reported the product bound; this is not defensible, because layers share failure causes — a stolen credential (layer 1) also defeats the rate profile expected by layer 5. We therefore report the general bounds. For any joint distribution:

$$\prod_{i=1}^{5} p_i \;\le\; Pr[X_1 \cap \ldots \cap X_5] \;\le\; \min_i p_i$$

where the left inequality holds when layer failures are non-negatively associated (the realistic case: compromising one layer never makes another *harder*) and the right inequality always holds. The system designer must therefore engineer the *minimum*, i.e., the strongest single layer, to the target ε, and treat the additional layers as reducing the gap between the bounds rather than multiplying safety. In DUALNET the strongest layer is by design the mandate mechanism (layer 3), whose failure probability under assumption 1 is the MAC forgery advantage — cryptographically negligible — so that the binding constraint on ε is the probability of Gate compromise itself, which is a governance quantity, not a cryptographic one. Figure 3 illustrates the product (lower) bound for intuition under three attacker profiles; it should be read as optimistic.

*Figure 3: Product (Independence) Lower Bound on Breach Probability*




### *8.2. Reduction summary*

The (q, ε)-separation of Definition 1 reduces as follows: a successful unsolicited delivery requires either (a) a forged return token — bounded by q times the MAC forgery advantage, negligible under assumption 1; (b) a forged personhood credential at threshold — excluded by assumption 2; or (c) compromise of G — excluded by assumption 3 and delegated to governance. Under the stated assumptions DUALNET achieves (q, ε)-separation with ε negligible in the security parameter; the honesty of the assumptions, particularly 3, is the true load-bearing element and is examined in Sections 9 and 10.

### *8.3. Covert channels*

Agents may attempt to signal human collaborators through patterns in legitimate traffic — the confinement problem [12]. The bandwidth of a timing/volume channel observable through G is bounded by

$$B_{covert} \leq log_2(N) \cdot (1/\Delta t),$$

where N is the number of distinguishable patterns G permits and Δt the granularity of observation. Gate-imposed random jitter reduces N; aggregation windows increase Δt. As noted after Definition 1, a human who decodes such a channel has initiated contact; the residual risk is therefore confined to colluding humans, which no network mechanism can prevent, and the design goal is to keep $B_{covert}$ below the rate at which structured payloads are practical.

### *8.4. Compromised ambassador*

If a user's ambassador agent is captured within AINet, it can return falsified information. Because all returns are type-checked and translated by G, falsification is constrained to content expressible under D: no executable payloads, no messages addressed to other users. The attack reduces to misinformation. Defense in depth dispatches k independent ambassadors and compares responses; consistent simultaneous compromise has probability $(p_{comp})^k$ under independent capture, and even under correlated capture the comparison forces the adversary to coordinate identical falsehoods across isolated agents, raising cost.

## 9. Governance and the Termination of the Dictator AI

The Dictator AI is a temporary AI authority that designs L-AI, configures G, and supervises initial deployment. After deployment it is terminated. Section 8 showed that the protocol's residual risk concentrates precisely here: ε is negligible *given* an honest Gate, and the Gate's honesty is an institutional property. This section examines who maintains it.

We identify three succession models:

**Constitutional model.** The Dictator AI's final configuration is treated as a constitution. Amendments require a supermajority of the Oversight Board and external ratification, analogous to constitutional amendment in democratic states. Advantage: stability and predictability. Disadvantage: inability to respond rapidly to novel attacks against L-AI or G.

**Sentinel model.** The Dictator AI is not terminated but placed in a non-modifying observational state: it can raise alarms to the Board but cannot alter L-AI, G, or the Capability Registry without explicit human consent. Advantage: ongoing adaptive supervision. Disadvantage: assumes that stasis-preserving suspension of an AI system is technically achievable and verifiable [5], which is itself an open problem.

**Phoenix model.** The Dictator AI is fully terminated; its specification (weights, training data, configuration) is archived under cryptographic seal with threshold-shared keys. In a crisis the Board may convene a limited revival under a narrowly scoped mandate, after which the instance is re-terminated. This most closely follows the CERN model of discrete supervised experiments [3]. Advantage: combines termination with recoverable capacity. Disadvantage: each revival is a potential capture event, and the revived instance may differ from the archived one in unaudited ways.

In all three models, trust transfers from the Dictator AI to the institution that manages its legacy. The long-term robustness of DUALNET therefore depends not on cryptography but on the design of an international oversight institution capable of resisting political capture while retaining technical competence over a system whose internal language no human can read. We regard the comparative institutional analysis — against precedents in arms control and international science governance — as essential future work rather than a solved input.

## 10. Limitations and Open Problems

We state plainly what this paper does not establish.

**No implementation.** DUALNET is an architecture and protocol design. We provide no proof-of-concept of the Translation Gate, no measured overhead for HTTP-H, and no empirical evaluation of the ingress classifier. These are the natural next steps and any deployment claim before them would be premature.

**No formal verification.** The protocol of Section 7 has not been verified in a tool such as ProVerif or Tamarin. The reduction argument of Section 8.2 is stated informally; mechanizing it would substantially strengthen the result and may surface attacks the prose argument misses.

**$C_H$ is heuristic.** The statistical human-language filter inherits the adversarial-example problem [14]. We mitigated this architecturally — the hard guarantee rests on the mandate mechanism, not on $C_H$ — but an adversary who defeats $C_H$ still gains a smuggling channel for human-readable content *within* AINet, which matters for the collusion scenarios of Section 8.3. Certified robustness for language classifiers is open.

**Lexical drift in L-AI.** The type system detects grammatical violations, but agents may evolve conventions that map existing well-typed terms onto human-referent meanings (an inner code). Detecting semantic drift without understanding the traffic is a fundamental tension: the Board audits a language it cannot read. Whether statistical audit suffices is unknown.

**The composite-actor boundary.** The human/agent dichotomy is increasingly false: humans operate with real-time AI assistance. DUALNET as specified would classify a human with an integrated assistant as a human, since the credential and mandate originate with the person; whether that is the right policy, and how to express graduated composite identities, is unresolved.

**Single gateway.** A single Translation Gate is a single point of failure and a scaling bottleneck. Distributing G across a threshold of operators is cryptographically straightforward for the MAC (threshold MACs exist) but institutionally untested for the translation and audit functions.

## 11. Conclusion

We have described DUALNET, a protocol for separating a network of human users from a network of autonomous AI agents while permitting one-directional query flow from humans to agents. Relative to prior isolation mechanisms, the contribution is asymmetric interactivity: data diodes restrict direction without sessions, anonymity networks hide identity without actor classes, and air gaps forbid interaction altogether. DUALNET combines proof-of-personhood, protocol incompatibility, a mediated and mandated translation gate, a constructed agent-side language, capability tokens, and behavioral detection, and we have argued that under explicit assumptions the construction achieves $(q, \varepsilon)$-separation with negligible $\varepsilon$, with the residual risk concentrated in the integrity of the gateway institution.

The technical construction is implementable with existing primitives. The harder problem is institutional: DUALNET requires a supervised AI authority to construct L-AI, an international board to govern the system after that authority is terminated, and political will to maintain the separation. Cryptography relocates trust; it does not abolish it.

Future work, in order of leverage: (i) mechanized verification of the mandated translation protocol; (ii) a proof-of-concept Gate with measured overhead; (iii) empirical study of ingress-classifier evasion; (iv) threshold distribution of the Gate; and (v) comparative institutional analysis of the three succession

models against historical precedents in international scientific and arms-control governance.

## References


[1] A. Acien, A. Morales, J. Fierrez, and R. Vera-Rodriguez, "BeCAPTCHA: Bot Detection in Smartphone Interaction Using Touchscreen Biometrics and Mobile Sensors," *IEEE Transactions on Information Forensics and Security*, 2021.

[2] T. Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," *arXiv preprint arXiv:2302.04761*, 2023.

[3] L. Maiani and L. Bonolis, "The LHC timeline: a personal recollection," *The European Physical Journal H*, vol. 42, no. 4, pp. 475–505, 2017.

[4] S. Adler et al., "Personhood Credentials: Artificial Intelligence and the Value of Privacy-Preserving Tools to Distinguish Who is Real Online," *arXiv preprint arXiv:2408.07892*, 2024.

[5] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[6] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using Hard AI Problems for Security," in *Advances in Cryptology — EUROCRYPT 2003*, Springer, 2003.

[7] J. R. Douceur, "The Sybil Attack," in *Peer-to-Peer Systems (IPTPS)*, Springer, 2002.

[8] D. Chaum, "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms," *Communications of the ACM*, vol. 24, no. 2, 1981.

[9] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The Second-Generation Onion Router," in *Proceedings of the 13th USENIX Security Symposium*, 2004.

[10] V. Costan and S. Devadas, "Intel SGX Explained," *IACR Cryptology ePrint Archive*, Report 2016/086, 2016.

[11] Worldcoin Foundation, "A New Identity and Financial Network," Whitepaper, 2023.

[12] B. W. Lampson, "A Note on the Confinement Problem," *Communications of the ACM*, vol. 16, no. 10, pp. 613–615, 1973.

[13] J. Camenisch and M. Stadler, "Proof Systems for General Statements about Discrete Logarithms," Technical Report TR 260, ETH Zürich, 1997.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[15] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[16] Anthropic, "Model Context Protocol Specification," 2024. [Online]. Available: https://modelcontextprotocol.io