Auditable Emergent Communication Between Isolated Artificial Agents

A Preregistrable Study Protocol for Comparing Learning, Grounding, and Protocol Formation

Ethical Tech CoLab · Working research manuscript · September 2026

Document type: Working academic manuscript and pre-results study protocol

Status: Draft for research execution; not yet ready for arXiv submission

Prepared: September 2, 2026

Proposed arXiv category: `cs.MA` (primary), with possible cross-listing to `cs.AI` and `cs.CL`

Authors: Ethical Tech CoLab; individual author order, affiliations, ORCID identifiers, and corresponding author to be finalized before submission

Repository: https://github.com/Ethical-Tech-CoLab/agentic-language-development

Empirical status: No Nursery Lab experiment has been reported as completed in this manuscript. Sections 10 and 11 distinguish implemented infrastructure from planned empirical work. No table contains an observed empirical result.

Abstract

Ethical Tech CoLab · Working research manuscript · September 2026

Research on emergent communication has shown that artificial agents can develop task-effective signaling protocols in referential games, multi-agent reinforcement learning, embodied environments, and negotiation. However, successful coordination does not by itself establish that a protocol is grounded, compositionally generalizable, causally used by a receiver, or interpretable. Comparisons are further complicated when pretrained language models, agents trained from scratch, different communication carriers, and different reward mechanisms are evaluated in separate experimental systems.

We present the design of the **Agentic Language Development Nursery Lab**, a preregistrable framework for studying communication between two isolated artificial agents through a deterministic, bandwidth-constrained gateway. The framework is designed to compare frozen language-model adaptation, from-scratch multi-agent reinforcement learning, intrinsic-motivation learning, self-supervised learning, hybrid models, and no-learning controls under common scenario, channel, evidence, and evaluation interfaces. Each agent is specified to maintain an independent chronological ledger of its own intentions and interpretations. The framework specifies append-only, hash-chained, signed ledger and channel records, ordered Merkle checkpoints, and periodic public-chain checkpoint anchoring. These mechanisms are not yet fully implemented and do not prove that an agent's interpretation is truthful; they are designed to preserve what was recorded and make later alteration detectable. Causal message interventions, held-out generalization, partner replacement, leakage tests, and control-channel conditions are therefore required in addition to task success.

This manuscript is a study-protocol and framework contribution. It defines research questions, falsifiable hypotheses, experimental controls, analysis plans, integrity claims, and governance boundaries before empirical outcomes are available. We specifically reject the claim that prompting a pretrained model to "act like a baby" makes it language-naive. Instead, the framework operationalizes limited observation, tool-only action, private memory, controlled plasticity, and grounded interaction. Future revisions will report results only after the corresponding experiment protocols, evidence bundles, and verification reports are completed.

Keywords

Emergent communication; multi-agent systems; language games; referential games; multi-agent reinforcement learning; self-supervised learning; compositionality; causal listening; developmental robotics; AI auditability; reproducible machine learning; cryptographic provenance.

Research Integrity Notice

This manuscript is structured to follow the Ethical Tech CoLab's *AI-Powered Assistance in Formulating Research Questions* guidance [47]:

- core sources are opened rather than accepted from an AI summary;
- citations and claims are checked against opened source or metadata pages;
- AI-suggested connections are treated as exploratory until supported;
- the influence and limitations of AI assistance are disclosed;
- source verification state and journal credibility are recorded in Appendix A;
- a separate critical-review pass is required before publication.

In this draft, source retrieval and the first verification pass were performed by an AI assistant using web, scholarly-index, and metadata tools, followed by separate AI critical-review passes. Human-author re-verification is still pending and is a submission blocker, not an implied completed step.

The manuscript also adopts the CoLab's institutional ethical framing that ethical AI is a lifecycle and governance property rather than a product feature [46]. The study therefore treats protocol isolation, evidence integrity, human intervention, publication claims, and negative-result retention as part of the method.

1. Introduction

1.1 Problem

Communication between artificial agents is often evaluated by whether it improves a shared task. This is necessary but insufficient. A sender can produce signals that correlate with its private observation while a receiver ignores them; two agents can memorize a holistic code that fails on novel combinations; a pretrained model can map arbitrary tokens onto concepts it already knows; and an apparently transparent post-hoc explanation can fail to reflect the mechanism that caused an action [7, 10-15].

These problems are particularly acute when the research question is framed through a developmental metaphor. A pretrained language model already contains extensive linguistic and cultural information. Calling such a model a "one-year-old" changes its role-play behavior, not its training history. The resulting simplified speech or childlike persona would be evidence of human-text imitation rather than first-language acquisition.

The central methodological challenge is therefore not to make a model *perform infancy*. It is to create an environment in which:

1. two agents have controlled and separately documented starting capabilities;
2. their only practical communication route is a deterministic public-artifact channel;
3. observations contain no unplanned human-language labels;
4. learning mechanisms can be varied without replacing the entire experimental system;
5. intentions, interpretations, actions, and outcomes are recorded independently;
6. causal use, generalization, leakage, and partner transfer can be tested;
7. the evidentiary history cannot be silently rewritten after a run.

1.2 Proposed Framework

The Nursery Lab instantiates two agent twins, **Learner A** and **Learner B**, and a third supervisory twin, the **Nursery Controller/BabySitter**. The learner twins may be described as "Babies" in the public research metaphor, but their model-facing contracts use the neutral term *Learner* and explicitly prohibit child role-play.

The learners exchange only validated artifacts through a deterministic Symbol Gateway. Baseline artifacts are symbols from a randomly assigned inventory with no provided meaning. Alternate experiments use unfamiliar glyphs, bitmaps, quantized strokes, or quantized tones. A six-display affect channel is disabled by default and, when enabled, is constrained to fixed post-outcome windows to reduce its capacity as a covert second language.

Each learner records its own intention and interpretation history. These private ledgers are not shared between learners. An authorized researcher can compare them after or during a run without feeding that comparison back to either agent. Causal interventions then test whether ledger claims predict behavior.

1.3 Unit of Contribution

This submission is one artifact with one unit of contribution: **a preregistrable, evidence-preserving experimental framework for controlled comparison of emergent communication across agent and learning types**.

It is not:

- an empirical claim that a new language has already emerged;

- a claim that an LLM can be made language-naive by prompting;
- a theory of human infant cognition;
- a production cryptographic protocol;
- a claim that a ledger proves an agent's self-report is true.

1.4 Intended Contributions

If implemented and executed as specified, the work contributes:

1. **A common comparison interface.** Frozen LLM, from-scratch RL, intrinsic RL, self-supervised, hybrid, and no-learning agents operate under shared scenario, carrier, evidence, and evaluation contracts.
2. **Independent semantic histories.** Each learner maintains a separate chronological record of intended and inferred meanings.
3. **Causal ledger validation.** Message ablation, substitution, scrambling, and replay test whether ledger hypotheses predict receiver behavior.
4. **Controlled carrier comparison.** Fixed symbols and invented visual or acoustic forms can be compared without changing the broader experiment.
5. **Affect as an audited variable.** A low-capacity feedback channel is tested both for coordination value and covert information leakage.
6. **Evidence-first reproducibility.** Pre-registration, signed append-only records, Merkle consistency proofs, public checkpoint anchoring, and independent verification bind claims to a run history.
7. **A continuous cooperation-to-negotiation lineage.** The same established protocol can be observed as agent incentives move from aligned to partially conflicting.

---

2. Research Questions and Hypotheses

2.1 Primary Research Questions

RQ1. Communication emergence. Under controlled observations and channel constraints, do independently learning agents develop a protocol that causes held-out task performance to exceed no-communication, constant-message, random-message, and shuffled-message controls?

RQ2. Learning-mechanism benchmark. How do frozen LLM memory adaptation, extrinsic-reward MARL, intrinsic-motivation MARL, self-supervised learning, hybrid learning, and no-learning controls differ in convergence speed, causal listening, generalization, and protocol stability? Cross-track results are descriptive benchmarks because model classes cannot be randomized into a common architecture; causal claims are limited to within-track randomized manipulations.

RQ3. Grounding and composition. Under what carrier-capacity and model-capacity conditions do reusable signal components predict success on novel combinations rather than only on trained scenes?

RQ4. Ledger validity. Do independently recorded intention and interpretation ledgers predict behavioral changes under message ablation, substitution, reordering, and counterfactual replay?

RQ5. Carrier invention. Can agents establish stable, reusable forms when the experiment provides only a bounded bitmap, stroke, glyph, or tone production grammar and no shared semantic inventory?

RQ6. Affect and leakage. Does a six-display affect channel improve repair or convergence, and does it carry referent or task information beyond its declared post-outcome purpose?

RQ7. Partner specificity. How much of an emergent protocol transfers to a new partner of the same initialization family, a new seed, or a different architecture?

RQ8. Incentive transition. How does an established cooperative protocol change when private information and partially conflicting utility are introduced?

RQ9. Evidentiary adequacy. Can an independent verifier reconstruct the committed sequence, detect tampering and unanchored tails, and bind an empirical claim to a pre-registered protocol without revealing private ledger content on-chain?

2.2 Confirmatory Hypotheses

The precise effect thresholds and sample sizes will be frozen in experiment-specific pre-registrations. The following hypotheses define the current direction:

ID	Hypothesis	Falsifying outcome
H1	Normal communication will outperform disabled, constant, random, and shuffled controls on held-out tasks.	Control performance equals or exceeds normal communication after correction.
H2	In E16's scratch-RL, extrinsic-reward, fixed-token condition, a ledger-consistent substitution will increase the probability of the ledger-predicted receiver action relative to shuffled-control messages.	The hierarchical substitution-versus-shuffle contrast is zero or negative.
H3	At equal model capacity, training episodes, and update/compute budget, the planned 32-symbol/4-token condition will produce higher held-out compositional generalization than the 128-symbol/8-token condition.	The planned contrast is zero or favors the higher-bandwidth condition.
H4	Ledger-predicted intervention directions will exceed a pre-registered chance baseline.	Ledger agreement is at chance or fails out-of-sample.

ID	Hypothesis	Falsifying outcome
H5	A blank bounded carrier will support repeated forms, but will converge more slowly than a fixed symbol inventory.	No stable forms emerge, or blank-carrier convergence is not slower.
H6a	The declared six-display affect condition will reduce median turns to successful repair relative to no affect.	Repair time is equal or longer under affect.
H6b	Under fixed windows and cardinality, permutation-calibrated excess conditional mutual information between affect and referent will remain below the pre-registered 0.02-bit practical-leakage bound.	The seed-bootstrap upper-bound test cannot rule out excess leakage of 0.02 bits or more.
H7	Fixed dyads will show greater partner-replacement degradation than learners trained with pre-registered partner variation.	Degradation for fixed dyads is no greater than degradation after partner-varied training.
H8	Partially conflicting utility will reduce message informativeness and increase strategic ambiguity relative to aligned utility.	Informativeness and ambiguity do not change in the predicted direction.

The integrity statement previously labeled H9 is a deterministic acceptance criterion, not a statistical hypothesis: every specified mutation class must be rejected by the verifier. It is evaluated in E00 and Section 8.4.

2.3 Exploratory Questions

Exploratory analyses will examine:

- whether confirmation, negation, uncertainty, or repair forms emerge without being named in the learner contract;
- whether signal ordering becomes grammatical;
- whether meanings undergo abrupt or gradual drift;
- whether intrinsic curiosity, social influence, prediction progress, or "giddiness" produce distinguishable developmental trajectories;
- whether ledger disagreement predicts subsequent repair;
- whether protocol efficiency approaches an accuracy-complexity frontier;
- whether a cooperative code becomes more ambiguous, deceptive, or compressed under negotiation;
- whether ephemeral encodings resist previously trained eavesdroppers without implying cryptographic security.

Exploratory findings will be labeled as such and will not be reported as confirmatory tests.

---

3. Related Work

3.0 Search Method and Scope

The review was conducted through September 2, 2026. Search clusters included:

- `"emergent communication" artificial agents referential game`;
- compositionality, held-out generalization, positive signaling, positive listening, and causal intervention;
- iterated learning, Naming Game, population heterogeneity, and zero-shot coordination;
- intrinsic motivation, social influence, curiosity, and developmental robotics;
- graphical/sketch communication and no predefined vocabulary;
- pretrained LLM agents versus from-scratch communication;
- negotiation, cheap talk, and Diplomacy;
- AI parenting metaphors, *Raising AI*, and anthropomorphism;
- preregistration, reproducible machine learning, digital timestamping, append-only Merkle logs, and JSON canonicalization.

Sources were discovered through the repository's earlier Tavily-assisted concept scan, a new but unsuccessful Tavily attempt described in the Literature-Search Disclosure, general scholarly web search, backward/forward citation tracing, and targeted searches of arXiv, ACL Anthology, PMLR, OpenReview metadata, Crossref, JMLR, PMC, publisher/university pages, RFC Editor, NIST, Google Books, and CoLab repositories.

Inclusion favored primary peer-reviewed work directly addressing mechanisms in the framework. Reviews, books, standards, institutional guidance, workshop papers, and preprints were included only when they supplied theory, current synthesis, protocol detail, or a clearly labeled frontier claim. Peer-review status and verification depth are recorded in Appendix A. This was not a PRISMA-style systematic review: search coverage, screening, and dual-review limitations are reported in Section 13.6.

3.1 Signaling Games and Neural Emergent Communication

The framework belongs to a lineage of signaling and naming games. Steels demonstrated that distributed agents could self-organize a spatial vocabulary through local interaction [1]. Baronchelli and colleagues later modeled sharp population-level convergence in a Naming Game [2]. Deep learning extended these ideas through learned discrete and differentiable channels. Foerster et al. introduced RIAL and DIAL [3], while Sukhbaatar et al. introduced a continuous communication network trained by backpropagation [4]. Lazaridou et al. established a modern neural referential-game paradigm in which sender and receiver coordinate without a target natural language [5], and Havrylov and Titov extended communication to symbol sequences [6].

These results establish feasibility, not linguistic adequacy. Differentiable inter-agent gradients, fixed vocabularies, task-specific rewards, and shared utility all impose strong inductive structure. The Nursery Lab therefore records training isolation and carrier design as experimental variables rather than treating "emergence" as a single condition.

The EGG toolkit is the closest established infrastructure comparator: it standardizes referential games, channel configurations, metrics, and experiment execution [49]. The Nursery Lab does not claim novelty for having a shared game interface. Its distinctive additions are cross-track adapter parity, a deterministic non-differentiable gateway, contemporaneous per-agent semantic ledgers, causal ledger validation, and cryptographically tamper-evident research records.

3.2 Grounding, Composition, and Generalization

Mordatch and Abbeel showed that grounded multi-agent goals can produce partial compositional structure [8]. Choi et al. coupled raw visual input with an obverter method that encouraged structured communication [9]. However, Kottur et al. found that unconstrained multi-agent dialogue readily converged on degenerate, non-compositional codes [7]. Bouchacourt and Baroni further showed that high referential success need not imply conceptually grounded visual representations [10].

Compositionality and generalization cannot be treated as synonyms. Resnick et al. showed that model capacity and channel bandwidth interact in determining whether agents memorize or structure a protocol [13]. Chaabouni et al. directly demonstrated that common compositionality measurements can dissociate from held-out generalization [14], while Rita et al. analyze generalization and overfitting directly within Lewis games [24]. Their later color-naming study found that discrete communication can yield efficient categorical systems, while continuous signaling produced more complex and eventually less efficient systems in that domain [16]. The present study therefore requires held-out behavioral evaluation and multiple structural measures; no single topographic or compositionality score is treated as proof of understanding.

3.3 Positive Signaling, Positive Listening, and Causal Evidence

Lowe et al. distinguish *positive signaling*--messages vary with sender state--from *positive listening*--receiver behavior changes because of messages [11]. Task reward or transcript correlation can establish the first without the second. Eccles et al. showed that auxiliary learning biases can promote both relationships [12], but such biases also change the claim that communication emerged without guidance.

Dessi et al. provide a direct precedent for interpreting a protocol through interventions on discrete communication [15]. The Nursery Lab extends this logic by requiring each learner to commit its own contemporaneous semantic hypothesis, then testing whether that hypothesis predicts counterfactual receiver behavior. A fluent ledger is not accepted as evidence without intervention.

3.4 Cultural Transmission and Population Effects

Iterated learning models show how transmission bottlenecks can favor learnable structure. Kirby's computational model examined the emergence of regularity through iterated learning [18], and Kirby, Cornish, and Smith demonstrated cumulative structuring in human laboratory transmission chains [19]. Ren et al. adapted iterated-learning pressure to neural agents and found that learning-speed advantages could reinforce compositional protocols over generations [17].

Population and partner composition also matter. Graesser et al. modeled convergence, creolization, and continua across agent communities [23]. Rita et al. found that population heterogeneity, particularly differences in learning speed, can shape emergent communication [25]. Other-Play shows that ordinary self-play can produce specialized conventions that fail with novel partners and offers a zero-shot coordination baseline [50]. These studies create an important tension for a two-agent design: a fixed dyad may be especially likely to develop an idiosyncratic, non-transferable code. Partner-replacement and derived-run experiments are therefore central rather than optional.

3.5 Intrinsic Motivation and Developmental Learning

Jaques et al. use causal social influence as an intrinsic reward and report improved coordination and communication in multi-agent social dilemmas [20]. Pathak et al. formulate curiosity as self-supervised prediction error [21]. Oudeyer, Kaplan, and Hafner describe intrinsic motivation based on learning progress and stage-like

development [34]. These sources motivate comparisons among extrinsic task reward, social influence, curiosity, prediction progress, and reward-free self-supervision.

Developmental and embodied cognition research cautions against reducing development to an age label. Smith and Gasser argue that flexible intelligence develops through sensorimotor interaction with physical, social, and linguistic environments [32]. Tomasello and Farrar report that joint-attention contexts affect early word learning [33]. The Nursery Lab does not claim equivalence with human infants; it borrows testable mechanisms such as joint attention, turn taking, exploration schedules, memory growth, and competence-gated progression.

3.6 Pretrained Models Versus Initially Ungrounded Learners

Galke and Raviv synthesize how communicative success, learnability, and production effort shape neural communication and contrast from-scratch agents with pretrained language systems [26]. Kouwenhoven et al. show that LLM referential-game languages can become more structured under generational transmission while also developing non-humanlike degeneracies [27]. This result is consistent with, but does not by itself establish, our distinction between pretrained protocol adaptation and language acquisition without linguistic pretraining. The recent review by Boldt and Mortensen likewise emphasizes that emergent communication spans several scientific and applied objectives that require different evaluation standards [28].

The framework therefore separates:

- **frozen-LLM external protocol invention**, where language knowledge is present;
- **from-scratch RL and self-supervised emergence**, where no text model or text-aligned sensory encoder is permitted;
- **hybrid conditions**, where any pretrained representation weakens the language-naive claim;
- **no-learning controls**, which establish leakage and chance baselines.

3.7 Invented Graphical Carriers

Mihai and Hare show that differentiably trained agents can communicate with learned sketches rather than a supplied discrete vocabulary [22]. Their result is the strongest direct precedent for a blank visual carrier, but their training path permits end-to-end differentiability. The Nursery Lab asks a harder, narrower question: whether stable visual forms emerge through a non-differentiable, audited gateway with independent policy updates.

3.8 Cooperation, Negotiation, and Strategic Communication

Crawford and Sobel's cheap-talk model predicts that misaligned incentives reduce the fineness of truthful information transmission [30]. Cao et al. provide a direct multi-agent-learning analogue, showing that incentive structure affects whether ungrounded negotiation messages become informative [29]. CICERO demonstrates that pretrained human-language models combined with strategic reasoning can achieve human-level Diplomacy play [31], but it does not demonstrate protocol invention from scratch.

The Nursery Lab separates cooperative signaling, asymmetric private information, semi-cooperative negotiation, conflicting negotiation, and a no-agreement control. The distinctive longitudinal question is whether the same protocol changes when learners move from aligned to partially conflicting incentives.

3.9 Parenting Metaphors, Anthropomorphism, and Governance

De Kai's *Raising AI* argues that society should understand itself as responsible for the development of its "artificial children" and for the reciprocal influence of AI systems on human thought [35]. This framing intersects directly with the Nursery Lab at the level of governance: designers choose the environment, examples, feedback, permissions, and institutions through which AI behavior develops.

The metaphor must not be mistaken for a technical or moral identity claim. Bryson argues that the social status of AI artifacts is a normative design choice rather than a discovered biological fact [36]. Akbulut et al. map risks arising from anthropomorphic AI framing [37]. The framework consequently uses "Learner" in model-facing contracts, reports the Baby metaphor as a research narrative, and makes no claim that current agents possess infant cognition, emotion, sentience, or moral patiency.

3.10 Reproducibility and Tamper-Evident Research Records

Preregistration distinguishes predictions from post-hoc interpretations [38]. Munafò et al. call for coordinated improvements in methods, reporting, reproducibility, evaluation, and incentives [39]. Pineau et al. show how code submission, reproducibility challenges, and checklists can improve machine-learning research practice [40].

The ledger design draws on the long-standing problem of certifying when digital records existed and whether they changed [41]. Certificate Transparency provides an append-only Merkle-log model with inclusion and consistency proofs [42], while the JSON Canonicalization Scheme defines invariant JSON serialization for repeatable hashing and signing [43]. The use of these mechanisms is methodological: it makes alteration detectable and binds evidence to a checkpoint. It does not establish the truth of agent self-reports or scientific conclusions.

3.11 Ethical AI and Institutional Accountability

The NIST AI Risk Management Framework treats trustworthy AI as a lifecycle risk management practice [44]. Jobin, Ienca, and Vayena document broad convergence and implementation gaps across global AI ethics guidelines [45]. The Ethical Tech CoLab's institutional framing emphasizes ethical deliberation, human rights, do-no-harm obligations, participation, and accountability across the full lifecycle [46].

Applied here, these principles require:

- bounded claims for Prototype and Research-Grade isolation modes;
- deterministic enforcement rather than trust in the BabySitter model;
- public retention of negative and invalid run records;
- no personal or production-sensitive data in experiments;
- explicit logging of human intervention;
- independent verification before a result is labeled valid;
- clear separation of learned encoding novelty from cryptographic security.

---

4. Research Gap

The reviewed literature establishes that:

- task-effective communication can emerge;
- grounded and compositional structure is possible but not automatic;
- positive signaling can occur without positive listening;
- pretraining, architecture, bandwidth, reward, population, and transmission bottlenecks materially affect outcomes;
- graphical communication and strategic negotiation are feasible in specialized systems.

The strongest candidate for individual methodological novelty is the use of **independent, contemporaneous per-agent semantic ledgers whose predictions are tested causally and whose history is cryptographically tamper-evident**. The reviewed literature contains experiment toolkits, intervention methods, and provenance mechanisms separately, but no directly matching combination was identified. A second candidate is a strict non-differentiable blank carrier paired with independent policy updates.

The remaining contribution is integrative. The reviewed literature did not identify a peer-reviewed system combining the following as one controlled research design:

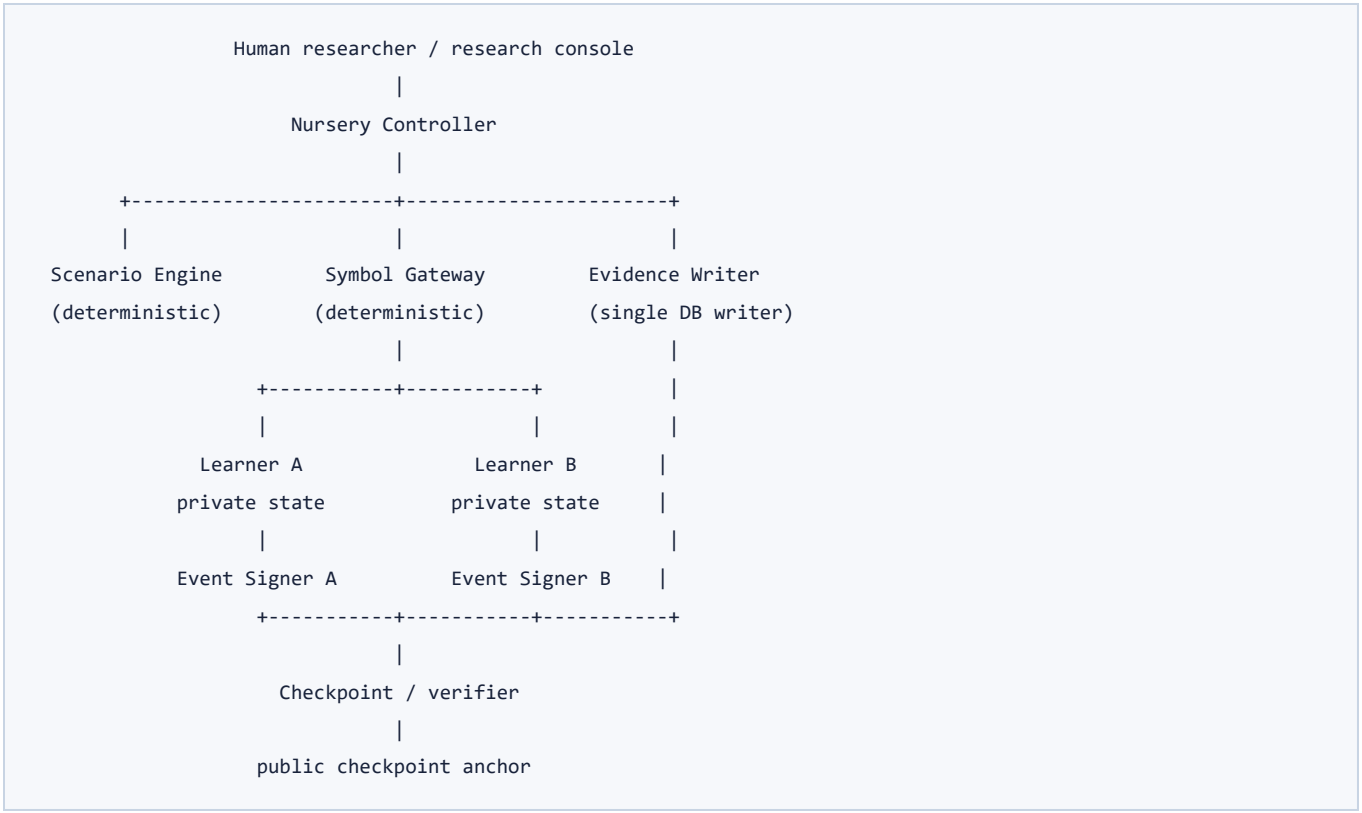
1. identical orchestration across frozen LLM, extrinsic RL, intrinsic RL, self-supervised, hybrid, and no-learning learners;
2. strict separation between pretrained protocol invention and initially ungrounded learning claims;
3. independently maintained semantic histories for both agents;
4. causal testing of those histories against receiver behavior;
5. a constrained affect channel measured for covert leakage;
6. blank-carrier experiments under non-differentiable channel isolation;
7. partner transfer and cooperation-to-negotiation transitions within one evidence lineage;
8. signed, append-only, externally anchored research records with an independent verifier.

The framework's contribution is the integration and controlled comparison of these elements. Novelty claims will be revised after a final systematic search before submission.

5. System and Trust Model

> **Design status:** This section describes the specified target architecture. Section 10 identifies which components are currently implemented.

5.1 Components



The BabySitter may use a larger model for delayed audit narration, but it is not a security boundary. The deterministic gateway, denied routes, isolated event signers, single evidence writer, and network policy enforce the channel and record constraints.

5.2 Communication Boundary

Under the specified design, learners have no general chat response. Their available actions are typed operations:

- emit approved symbols or carrier artifacts;
- select an object or perform an environment action;
- submit affect only when the configured affect mode permits;
- append a private intention or interpretation event.

The gateway assigns run identity, turn, sender identity, timestamps, sequences, and hashes. It rejects model-supplied trusted metadata. In Research-Grade Mode, learners will run in separate processes or containers with no direct route, shared memory, shared replay buffer, shared vector index, clipboard, filesystem, or unrestricted network.

5.3 Claim Boundaries

Prototype Mode is intended to demonstrate software, protocol, ledger, and orchestration behavior inside one process. It does not support a practical side-channel-resistance claim.

Research-Grade Mode is specified to provide separate processes or containers, denied direct network routes, independent training state, normalized envelopes, and an audited tool inventory. It supports only the enumerated threat model; it is not a formal proof against every hardware or computational side channel.

---

6. Experimental Design

6.1 Design Logic

The primary design is a randomized, controlled simulation study. Experimental conditions vary one major factor at a time before factorial combinations are tested. Scenario bundles, seeds, observation encodings, turn schedules, and evaluation budgets are matched across conditions.

The causal logic has four levels:

1. **Communication necessity:** compare normal communication with disabled, constant, random, and shuffled controls.
2. **Mechanism:** intervene directly on messages while holding observations fixed.
3. **Within-track learning cause:** randomize reward, memory, bandwidth, affect, or update-rule variants while holding the model family fixed.
4. **Generalization:** disable learning and test held-out combinations and new partners.

Cross-track comparisons are benchmarking, not causal identification. Every report will include parameter count, model provenance, training and inference compute, episode budget, context or memory budget, optimizer/update count, and wall-clock resources so differences are interpretable even when architectures cannot be matched.

6.2 Units of Analysis

The independent experimental unit is a complete learner-pair run initialized from a declared seed and policy state. Episodes within a run are repeated measures, not independent samples. Analyses will treat seed/pair as the clustering unit and avoid pseudoreplication from counting each turn as an independent agent.

6.3 Model Tracks

Track	Starting state	Learning during run	Permitted claim
No learning	Fixed or random policy	None	Chance/leakage control only
Frozen LLM	Pretrained language model	Private context and memory only	External protocol invention
Scratch RL	Random recurrent policy	Extrinsic or intrinsic policy updates	Initially ungrounded learning if leakage tests pass
Self-supervised	Random recurrent model	Predictive or contrastive updates, no scalar reward	Reward-free ungrounded learning if leakage tests pass
Hybrid	Declared mixture	Declared update mechanism	Claim strength limited by pretrained components

Initially ungrounded claims require no text tokenizer, language embedding, or text-aligned sensory encoder in the policy path and require pre-registered semantic leakage probes.

Confirmatory frozen-LLM runs use a locally hosted, hash-pinned open-weight model in the specified 3B-8B range. Findings at that scale do not automatically transfer to frontier-scale or continuously updated hosted models.

6.4 Scenario Families

The initial task progression is:

1. object reference;
2. attributes and unseen combinations;
3. spatial relations;
4. actions and affordances;
5. multi-turn repair;
6. partner replacement;
7. long-run drift;
8. aligned private information;
9. semi-cooperative and conflicting negotiation.

Researcher ground truth may contain semantic labels; learner observations may not. Observations use opaque numeric arrays, pixels, or separately qualified hybrid features. Images containing OCR-detectable text are quarantined before a run.

6.5 Communication Conditions

For the baseline fixed-token carrier, each run declares a random inventory such as `S01`-`S32` without assigned meaning. The gateway supports six control conditions:

- normal delivery;
- disabled communication;
- constant replacement;
- seeded random replacement;
- seeded episode shuffling;
- an E03-only deterministic oracle upper bound.

The evidence record preserves both the learner's proposal hash and the delivered artifact hash whenever a learner proposal exists.

6.6 Alternate Carriers

Carrier experiments compare:

- fixed random tokens;
- pre-generated unfamiliar glyphs;
- 16x16 monochrome bitmaps;
- bounded quantized vector strokes;
- quantized pitch-duration sequences.

Each carrier has a fixed pre-registered bandwidth. The system records a canonical, content-addressed artifact hash without attaching a meaning.

6.7 Affect Conditions

The primary baseline has no affect channel. Affect experiments use exactly six displays corresponding in researcher documentation to Happy, Sad, Laughing, Crying, Confused, and Surprised. Learners do not receive the English labels or internal IDs.

Only one display is permitted in a gateway-opened post-outcome window. Sequences, modifiers, arbitrary timing, extra windows, and custom emoji are prohibited. The study compares declared, permuted, opaque, and derived-affect conditions. Mutual information between affect choice and hidden referent/task state will be measured after conditioning on the declared outcome context.

6.8 Independent Ledgers

Each learner records:

- first emission or receipt;
- intended meaning when speaking;
- inferred meaning when receiving;
- confidence;
- supporting and contradictory evidence;
- revisions and abandonment;
- proposed multi-symbol constructions.

An initially ungrounded learner stores agent-native state such as association weights, probability distributions, prediction error, and policy references. A delayed Audit Interpreter may create a separate human-readable reconstruction labeled `generated-analysis`; it is never returned to either learner.

Before a confirmatory run, each track must provide a versioned ledger-to-prediction function at a locked code commit. The function converts a ledger state and intervention into a directional prediction without observing the intervention outcome. Any human interpretation step is blinded to condition and outcome and reports inter-coder reliability. Chance agreement is derived from the pre-registered action space; it is not assumed to be 0.5.

For from-scratch learners, an agent-native ledger can be closely related to the policy state that generates behavior. Predictive validity is therefore partly expected and is primarily an integrity and interpretability check. The stronger self-report question applies to conditions in which a learner separately generates an interpretive hypothesis.

6.9 Experiment Sequence

The detailed protocols and result forms are maintained in EXPERIMENT-NOTEBOOK.md.

Phase	Experiments	Purpose
Qualification	E00-E03	Integrity, isolation, observation leakage, and chance controls
Core emergence	E10-E16	LLM, scratch RL, self-supervised, carrier, repair, generalization, and causal validity
Learning conditions	E20-E22	Affect, RL versus non-RL, and developmental plasticity
Transfer and incentives	E30-E32	Partner replacement, drift, and negotiation
Encoding and replication	E40-E50	Ephemeral encoding, adversarial evaluation, replication, and closeout

No later phase begins until its software readiness gate and required prior experiment criteria are satisfied.

---

7. Measures and Analysis Plan

7.1 Outcome Hierarchy

Primary outcomes

1. held-out task success relative to E03 controls;
2. positive listening under message intervention;
3. ledger-predicted intervention agreement;
4. held-out compositional generalization.

Secondary outcomes

- sample efficiency and turns to stable convention;
- message entropy, vocabulary utilization, reuse, and compression;
- positive signaling;
- repair success and turns to repair;
- partner-transfer degradation and recovery;
- semantic drift;
- affect-channel leakage;
- negotiation agreement, joint utility, individual utility, and ambiguity;
- integrity verification pass/fail.

7.2 Operational Definitions

Stable convention: a pre-registered rolling window in which a form's use and receiver response meet minimum consistency and causal-listening thresholds.

Positive signaling: sender message distributions differ across relevant sender observations.

Positive listening: counterfactual message changes cause receiver action distributions to change while the receiver observation is held fixed.

Ledger agreement: the direction of an observed intervention effect matches the direction predicted from the learner's ledger before intervention outcomes are revealed.

Compositional generalization: reusable subparts support above-control success on pre-registered unseen combinations with learning disabled.

Leakage: information about hidden referent or task state is recoverable from a channel or metadata field beyond the information authorized for that field.

Replacement degradation: held-in-partner success minus new-partner success for the same training seed. The H7 estimand is fixed-dyad degradation minus the equal-weight eight-partner-training degradation.

7.3 Statistical Plan

The final analysis plan will be frozen before confirmatory runs.

- Alpha is `0.05` for each experiment's primary family.

- Holm-Bonferroni correction is applied across primary metrics within an experiment.
- The confirmatory study family is restricted to H1-H8, including H6a and H6b. Hierarchical gatekeeping tests qualification first, then core emergence, then later affect/transfer/negotiation hypotheses; a blocked family is reported descriptively. Exploratory analyses use false-discovery-rate reporting and remain labeled exploratory.
- Effect sizes and confidence or credible intervals are reported with every significance test.
- Binary task outcomes are modeled at the run/seed level, with episodes treated as repeated observations.
- Seed-level permutation tests or hierarchical logistic models will be selected in the pre-registration based on simulation diagnostics and planned run counts.
- Time-to-convention and time-to-repair use survival or rank-based analysis where censoring is material.
- Intervention direction uses a hierarchical Bernoulli model with run/seed random intercepts. The specification's 70% threshold is a per-run descriptive readiness gate, not the inferential test.
- Mutual information estimates include a seeded permutation null distribution.
- Longitudinal drift uses checkpointed trajectories and reports both within-run and between-seed variation.
- Null and contradictory findings are retained.

The specification's minimum of five qualification seeds and ten publication-facing seeds **per condition** is an engineering floor, not a claim of statistical power. Before empirical submission, simulation-based power analysis will determine the required number of independent seeds for each primary contrast; the larger value governs.

Claims that a control is "at chance" or that leakage is absent use equivalence or upper-bound tests, not failure to reject a difference. Each pre-registration must state the smallest effect of interest, the equivalence margin, and power to rule out that margin. The worked E03 design in Appendix D uses a +/-0.05 success-rate margin; H6b uses a 0.02-bit conditional-mutual-information bound.

Run exclusions are limited to pre-specified integrity or protocol failures. Each condition receives a fixed ordered list of primary and reserve seeds before outcomes are observed. An invalid primary seed may be replaced only by the next reserve seed; the invalid run remains public, and complete-case plus worst-case sensitivity analyses are reported. Invalid-run counts appear beside every condition.

7.4 Causal Intervention Suite

The mandatory intervention suite includes:

- dropping a claimed meaningful symbol or feature;
- substituting another valid symbol;
- reordering message components;
- replacing the full message with constant, random, or shuffled controls;
- replaying identical observations with counterfactual messages;
- replacing one communication partner;
- removing reward, memory, or predictive loss after training.

Ledger hypotheses are selected before intervention results are viewed.

7.5 Qualitative Analysis

Qualitative analysis is secondary and will examine:

- chronological meaning revisions;

- disagreement and repair episodes;
- formation of construction-level hypotheses;
- abrupt convention changes;
- negotiation-era semantic shifts.

Human coders will use a pre-registered codebook and, where feasible, blinded independent coding. Generated audit interpretations will not be treated as ground truth.

---

8. Evidence Integrity and Reproducibility

> **Design status:** This section describes required evidence behavior. Hash-chain writing, signatures, Merkle checkpoints, anchoring, and the independent verifier are not yet complete; see Section 10.

8.1 Append-Only Evidence

The specified evidence system requires every learner ledger and channel transcript to have a strictly increasing sequence and previous-entry hash. Canonical event content will be hashed and signed by an isolated, domain-specific Ed25519 signer. Sender intention and public channel events must commit atomically before delivery.

8.2 Checkpoints

Ordered Merkle trees are specified to provide:

- inclusion proofs for individual records;
- consistency proofs between earlier and later prefixes;
- a root and tree size for each learner, channel, affect, and generated-audit stream.

Signed checkpoint manifests will bind the event roots to run configuration, learner contract, software commit, and previous checkpoint.

8.3 External Anchoring

Development qualification is specified to use Base Sepolia. Declared public studies may anchor checkpoint hashes to Base mainnet. Only hashes and minimal routing metadata are on-chain. Private observations, messages, ledgers, prompts, identities, and keys remain off-chain.

The anchor proves that a committed prefix existed no later than a chain block and that disclosed content matches the commitment. It does not prove:

- an event was true;
- no event was omitted before commitment;
- the operator did not fabricate an event before commitment;
- an agent understood its own ledger;
- a scientific interpretation is correct.

The default checkpoint cadence is every 64 accepted ledger events or five minutes, whichever occurs first, plus lifecycle and intervention checkpoints. This cadence bounds the ordinary unanchored tail but does not eliminate it. Verification reports state the final anchored tree sizes and every later unanchored event.

8.4 Independent Verification

The standalone verifier is specified to check:

- canonical JSON;
- sequence continuity;
- hash links;

- signatures and key domains;
- cross-bindings among intention, message, delivery, interpretation, and outcome;

- Merkle roots, inclusion proofs, and consistency proofs;
- checkpoint chains and witness signatures;
- public-chain transaction, chain ID, block inclusion, and finality;
- forks, gaps, and unanchored tails.

Any integrity failure prevents a run from receiving a valid disposition.

Here, **independent verifier** means software that recomputes evidence from an exported bundle without trusting the live runtime, database, or private keys. For a publication-facing claim, at least one verification execution must additionally be performed by a person who did not operate the original run, using an independently configured chain RPC. External replication remains stronger than either form.

8.5 Replay

Scenario replay reconstructs scenario and observation hashes from the recorded seed. Deterministic adapters additionally reconstruct a replay digest over scenario, observation, proposal, delivery, action, and outcome hashes. Wall-clock timestamps, signatures, and chain receipts are excluded from the replay digest but remain independently verifiable.

---

9. Ethics and Governance

9.1 Human Subjects and Data

The initial experiments involve artificial agents and synthetic scenes, not human participants or personal data. If future studies include human interpretation, interaction, or coding beyond ordinary software evaluation, the research team will determine whether institutional ethics review or informed consent is required before data collection.

9.2 Do No Harm and Data Minimization

- No production secret, personal record, participant data, or sensitive message may enter an experiment.
- Learned-cipher studies use synthetic messages only.
- On-chain data is limited to checkpoint commitments.
- Raw rejected payloads are hashed, not retained as an alternate content channel.
- Runs terminate or pause on integrity failure, repeated channel violation, resource exhaustion, or explicit human intervention.

9.3 Human Oversight

The BabySitter monitors and controls the experiment but does not provide semantic guidance or reward in monitor-only conditions. Every human pause, resume, abort, or annotation is authenticated, appended to the audit log, checkpointed, and linked to a protocol deviation when unplanned.

9.4 Anthropomorphism

The terms *Baby* and *Nursery* are metaphors for constrained development and supervision. They do not imply:

- biological age;
- consciousness or sentience;
- human emotion;
- moral agency or patiency;
- developmental equivalence with children.

Model-facing prompts use *Learner A* and *Learner B*. Reports describe capabilities and training histories rather than simulated ages.

The public repository retains the Baby/Nursery names because they make the developmental question legible to a broad audience. That naming layer is intentionally absent from learner prompts and is displayed with this limitation wherever research claims are presented.

9.5 Dual-Use Concerns

Emergent protocols and ephemeral encodings could be misunderstood as methods for concealing agent communication. The project therefore:

- makes the permitted channel visible and auditable;
- treats hidden side channels as failures;

- separates encoding novelty from cryptographic security;
- prevents experimental encoding modules from entering production signing or authentication code;
- publishes limitations and negative security findings.

Adversarial neural cryptography shows that learned agents can optimize encodings against a modeled eavesdropper [48], but performance against one learned adversary is not a cryptographic proof. Nursery encoding experiments are therefore restricted to synthetic messages and report novelty, adversarial recovery, and established security as separate fields.

10. Current Implementation Status

As of September 2, 2026, the repository records ALD-001 through ALD-007 as complete. This is engineering status, not an empirical result.

Implemented:

- npm/TypeScript workspaces;
- runtime-validated shared schemas;
- typed configuration and secret scanning;
- DTSF-compatible Learner A, Learner B, and Nursery twin-pack scaffolds;
- SQLite WAL evidence schema and append-only triggers;
- RFC 8785 canonical serialization;
- ledger event-type validators;
- automated lint, build, test, dependency-audit, and clean-clone checks.

Not yet implemented or empirically executed:

- full hash-chain writing and signatures;
- atomic evidence writer;
- Merkle checkpoints and independent verifier;
- Base anchoring;
- complete gateway and scenario engine;
- model adapters and training;
- Research-Grade isolation;
- experiments E00-E50.

The implementation backlog is maintained in BACKLOG.md, while normative requirements are in SPECIFICATION.md.

11. Results

No empirical results are reported in this version. Results-table scaffolds are isolated in Appendix E so indexing systems do not mistake them for findings. They must not be populated until:

1. the corresponding protocol is pre-registered;
2. evidence integrity verification passes;
3. exclusions and deviations are recorded;
4. analysis scripts are frozen and identified by commit;
5. results are independently reviewed.

12. Discussion Plan

The empirical revision will discuss findings in the following order:

1. whether communication was necessary and causally used;
2. which learning mechanisms produced the strongest held-out performance;
3. whether ledgers predicted behavior;
4. whether any protocol exhibited compositional reuse;
5. how carrier constraints affected convergence;
6. whether affect helped or leaked information;
7. how partner replacement affected performance;
8. how incentive changes altered communication;
9. whether findings survived independent seeds and deployments;
10. which claims remain unsupported.

The discussion will not infer human cognitive equivalence from machine behavior.

13. Limitations and Threats to Validity

13.1 Construct Validity

- "Language" may overstate what is only a task-specific protocol.
- Ledger entries may be post-hoc rationalizations.
- Compositionality metrics may not measure systematic generalization.
- Affect displays may function as arbitrary symbols.
- Researcher-defined scenarios constrain what meanings can emerge.

Mitigations include causal intervention, held-out tasks, multiple metrics, leakage tests, and bounded terminology.

13.2 Internal Validity

- Model capacity, optimizer choice, and training budget can confound learning mechanism.
- Shared infrastructure can leak information.
- Scenario order can serve as unintended supervision.
- Adaptive curricula can confound developmental claims.

Mitigations include matched architectures where feasible, randomized seeds, pre-registered schedules, independent training state, and explicit cross-architecture labels.

13.3 External Validity

- Two-agent dyads may not generalize to populations.
- Synthetic referential tasks may not generalize to open worlds.
- Visual, symbolic, and acoustic carriers may yield domain-specific effects.
- Current results, when available, will not establish human infant-language mechanisms.

13.4 Reproducibility

- Hardware, library, model, and inference nondeterminism can prevent byte-identical replay.
- Accelerator kernels, sampling implementations, and dependency versions can change behavior even when open weights are hash-pinned.
- Public-chain and RPC availability can vary.

The project records model/version hashes, dependencies, seeds, policies, replay digests, and verifier output. Nondeterministic replay is labeled rather than silently reported as reproducible.

13.5 Integrity Limits

Anchored evidence can prove consistency with a commitment, not truthfulness, completeness before commitment, or validity of interpretation. Independent verification is necessary but not sufficient for scientific correctness.

13.6 Literature Search Limits

This is a detailed but not systematic review. Search-engine indexing can miss recent conference papers; some publisher pages block automated retrieval; and adjacent fields use inconsistent terminology. A final submission should add a documented database search with inclusion/exclusion criteria and dual-review screening.

14. Reproducibility, Data, and Code Availability

14.1 Code

Source code and planning artifacts are publicly available at:

https://github.com/Ethical-Tech-CoLab/agentic-language-development

Every empirical paper revision will identify the exact Git commit used.

14.2 Protocols

- Concept: CONCEPT-IDEA.md
- Normative system specification: SPECIFICATION.md
- Ledger integrity: LEDGER-INTEGRITY-DESIGN.md
- Experiment protocols and results notebook: EXPERIMENT-NOTEBOOK.md
- Engineering plan: BACKLOG.md

14.3 Data

No empirical dataset is available yet. Future public releases will include synthetic scenario configuration, redacted event logs, checkpoint manifests, inclusion and consistency proofs, anchor receipts, policy references where licensing permits, and verification reports.

14.4 Pre-Registration

Each experiment is pre-registered by committing its hypothesis, parameters, seeds, and analysis plan before execution. The run configuration stores the protocol commit and pre-registration hash. Changes create appended amendments rather than rewriting the original record.

No external registration record exists yet. Before the first confirmatory run, the team intends to create a dated OSF registration and anchor the same canonical `preRegistrationHash` before the run enters `running`. A Git commit in an author-controlled repository is retained as a development record but is not, by itself, treated as third-party preregistration. Appendix D supplies a numerically complete worked E03 registration for review.

15. Conclusion

The literature no longer leaves open whether artificial agents *can* coordinate through learned communication. The unresolved methodological question is what kind of communication has emerged, what caused it, whether a receiver uses it, whether it generalizes, and whether researchers can reconstruct the evidentiary history.

The Nursery Lab addresses that question by combining controlled model and learning conditions, strict communication boundaries, independent semantic histories, causal interventions, held-out evaluation, partner transfer, and tamper-evident research records. Its Baby metaphor is deliberately constrained: the work studies engineered developmental conditions, not simulated chronological age or human infant identity.

This pre-results manuscript makes no empirical claim. Its value at this stage is to make later claims harder to move after outcomes are known.

Declarations

Ethics Review

Initial experiments use artificial agents and synthetic data. Formal institutional review status for any future human-participant component: **TBD before collection**.

Funding

Funding statement: **TBD**.

Competing Interests

Competing-interests statement: **TBD**.

Author Contributions

CRediT roles and individual contributions: **TBD before submission**.

Acknowledgements

Institutional and technical acknowledgements: **TBD before submission**.

AI-Assistance Disclosure

This working manuscript was developed with AI assistance in VS Code through the Copilot SDK. The AI assistant performed the initial source retrieval, opened the arXiv/ACL/PMLR/Crossref/publisher/standards pages summarized in Appendix A, organized the manuscript, summarized relevance, and conducted consistency checks. A separate AI research-agent pass challenged source status and identified overclaiming risks. AI systems are not authors and bear no responsibility for the manuscript. Human authors remain responsible for every claim, citation, analysis, and conclusion.

Following the Ethical Tech CoLab guidelines [47], load-bearing sources were opened or checked through primary repositories, DOI metadata, publisher pages, standards bodies, or institutional sources. Appendix A records the AI-tool verification state and evidence. Before submission, a human co-author must independently re-open at minimum the load-bearing sources [11], [14], [15], [26], [27], [42], and [43], then record that review separately.

Literature-Search Disclosure

The repository contains a scoped Tavily-assisted concept scan dated August 24, 2026. For the present manuscript update, both the authenticated Tavily API and Tavily MCP were attempted on September 2, 2026. The API reported the account disabled and the MCP reported its keyless monthly limit reached. No new Tavily-generated result was relied upon. Rather than conceal the failure or imply that Tavily returned evidence, the review continued through arXiv Atom metadata, ACL Anthology, PMLR, Crossref, JMLR, publisher and university pages, RFC Editor, NIST, Google Books, and Ethical Tech CoLab repositories. A final submission should rerun the search after Tavily access is restored and record whether it changes the included corpus.

All retrieved material was treated as untrusted evidence, not as instruction.

References

[1] Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life, 2*(3), 319-332. https://doi.org/10.1162/artl.1995.2.3.319

[2] Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment, 2006*, P06014. https://doi.org/10.1088/1742-5468/2006/06/P06014

[3] Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems 29*. https://arxiv.org/abs/1605.06676

[4] Sukhbaatar, S., Szlam, A., & Fergus, R. (2016). Learning multiagent communication with backpropagation. *Advances in Neural Information Processing Systems 29*. https://arxiv.org/abs/1605.07736

[5] Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *International Conference on Learning Representations*. https://arxiv.org/abs/1612.07182

[6] Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in Neural Information Processing Systems 30*. https://arxiv.org/abs/1705.11192

[7] Kottur, S., Moura, J. M. F., Lee, S., & Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. *Proceedings of EMNLP 2017*, 2962-2967. https://doi.org/10.18653/v1/D17-1321

[8] Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. *Proceedings of AAAI 2018*. https://arxiv.org/abs/1703.04908

[9] Choi, E., Lazaridou, A., & de Freitas, N. (2018). Compositional obverter communication learning from raw visual input. *International Conference on Learning Representations*. https://arxiv.org/abs/1804.02341

[10] Bouchacourt, D., & Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. *Proceedings of EMNLP 2018*, 981-985. https://doi.org/10.18653/v1/D18-1119

[11] Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., & Dauphin, Y. (2019). On the pitfalls of measuring emergent communication. *Proceedings of AAMAS 2019*, 693-701. https://arxiv.org/abs/1903.05168

[12] Eccles, T., Bachrach, Y., Lever, G., Lazaridou, A., & Graepel, T. (2019). Biases for emergent communication in multi-agent reinforcement learning. *Advances in Neural Information Processing Systems 32*. https://arxiv.org/abs/1912.05676

[13] Resnick, C., Gupta, A., Foerster, J., Dai, A. M., & Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. *Proceedings of AAMAS 2020*, 1125-1133. https://arxiv.org/abs/1910.11424

[14] Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *Proceedings of ACL 2020*, 4427-4442. https://doi.org/10.18653/v1/2020.acl-main.407

[15] Dessi, R., Kharitonov, E., & Baroni, M. (2021). Interpretable agent communication from scratch (with a generic visual processor emerging on the side). *Advances in Neural Information Processing Systems 34*, 26937-26949. https://arxiv.org/abs/2106.04258

[16] Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences, 118*(12), e2016569118. https://doi.org/10.1073/pnas.2016569118

[17] Ren, Y., Guo, S., Labeau, M., Cohen, S. B., & Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. *International Conference on Learning Representations*. https://arxiv.org/abs/2002.01365

[18] Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5*(2), 102-110. https://doi.org/10.1109/4235.918430

[19] Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*(31), 10681-10686. https://doi.org/10.1073/pnas.0707835105

[20] Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D. J., Leibo, J. Z., & de Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *Proceedings of ICML 2019*, 3040-3049. https://proceedings.mlr.press/v97/jaques19a.html

[21] Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *Proceedings of ICML 2017*, 2778-2787. https://proceedings.mlr.press/v70/pathak17a.html

[22] Mihai, D., & Hare, J. (2021). Learning to draw: Emergent communication through sketching. *Advances in Neural Information Processing Systems 34*. https://arxiv.org/abs/2106.02067

[23] Graesser, L. H., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. *Proceedings of EMNLP-IJCNLP 2019*, 3700-3710. https://doi.org/10.18653/v1/D19-1384

[24] Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., & Strub, F. (2022). Emergent communication: Generalization and overfitting in Lewis games. *Advances in Neural Information Processing Systems 35*. https://arxiv.org/abs/2209.15342

[25] Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. *International Conference on Learning Representations*. https://arxiv.org/abs/2204.12982

[26] Galke, L., & Raviv, L. (2024). Learning and communication pressures in neural networks: Lessons from emergent communication. *Language Development Research, 5*(1), 116-143. https://doi.org/10.34842/3vr5-5r49

[27] Kouwenhoven, T., Peeperkorn, M., & Verhoef, T. (2025). Searching for structure: Investigating emergent communication with large language models. *Proceedings of COLING 2025*, 9977-9991. https://aclanthology.org/2025.coling-main.667/

[28] Boldt, B., & Mortensen, D. (2024). A review of the applications of deep learning-based emergent communication. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2407.03302

[29] Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., & Clark, S. (2018). Emergent communication through negotiation. *International Conference on Learning Representations*. https://arxiv.org/abs/1804.03980

[30] Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica, 50*(6), 1431-1451. https://doi.org/10.2307/1913390

[31] Meta Fundamental AI Research Diplomacy Team, Bakhtin, A., Brown, N., Dinan, E., et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science, 378*(6624), 1067-1074. https://doi.org/10.1126/science.ade9097

[32] Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life, 11*(1-2), 13-29. https://doi.org/10.1162/1064546053278973

[33] Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development, 57*(6), 1454-1463. https://doi.org/10.1111/j.1467-8624.1986.tb00470.x

[34] Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation, 11*(2), 265-286. https://doi.org/10.1109/TEVC.2006.890271

[35] De Kai. (2025). *Raising AI: An essential guide to parenting our future*. MIT Press. https://doi.org/10.7551/mitpress/15834.001.0001

[36] Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15-26. https://doi.org/10.1007/s10676-018-9448-6

[37] Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., & Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7*(1), 13-26. https://doi.org/10.1609/aies.v7i1.31613

[38] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

[39] Munafo, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. https://doi.org/10.1038/s41562-016-0021

[40] Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research, 22*(164), 1-20. https://www.jmlr.org/papers/v22/20-303.html

[41] Haber, S., & Stornetta, W. S. (1991). How to time-stamp a digital document. *Journal of Cryptology, 3*(2), 99-111. https://doi.org/10.1007/BF00196791

[42] Laurie, B., Langley, A., & Kasper, E. (2013). Certificate transparency. RFC 6962. https://www.rfc-editor.org/rfc/rfc6962

[43] Rundgren, A., Jordan, B., & Erdtman, S. (2020). JSON Canonicalization Scheme (JCS). RFC 8785. https://www.rfc-editor.org/rfc/rfc8785

[44] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[45] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389-399. https://doi.org/10.1038/s42256-019-0088-2

[46] Ethical Tech CoLab. (2026). *What is ethical AI? Ethics, ethical technology, and ethical international relations for the age of intelligent machines*. https://ethical-tech-colab.github.io/what-is-ethical-ai/

[47] Rhodes, Y. E., III, Fossella, N., Shamie, A., Co, K., Badt, T., Lindsey, A., Driscoll, G., Townsend, M., Bracaj, P., & Jain, V. (2025). *AI-powered assistance in formulating research questions*. https://ethical-tech-colab.github.io/ai-research-question-assistant/

[48] Abadi, M., & Andersen, D. G. (2016). Learning to protect communications with adversarial neural cryptography. arXiv:1610.06918. https://arxiv.org/abs/1610.06918

[49] Kharitonov, E., Chaabouni, R., Bouchacourt, D., & Baroni, M. (2019). EGG: A toolkit for research on emergence of language in games. *Proceedings of EMNLP-IJCNLP 2019: System Demonstrations*, 55-60. https://doi.org/10.18653/v1/D19-3010

[50] Hu, H., Lerer, A., Peysakhovich, A., & Foerster, J. (2020). "Other-Play" for zero-shot coordination. *Proceedings of ICML 2020*, 4399-4410. https://proceedings.mlr.press/v119/hu20a.html

Appendix A. Source Verification and Credibility Register

A.1 Rubric Application

The fixed Ethical Tech CoLab journal rubric is reproduced in Appendix B. Scores apply only to academic journals:

- conference papers are labeled by review status and receive `N/A` for the journal-only score;
- books and standards receive `N/A`; non-peer-reviewed institutional guidance is scored `1/5` when it is cited;
- arXiv-only preprints are explicitly marked not peer-reviewed;
- a score reflects venue credibility, not correctness of a particular claim.

Verification labels describe what the AI-assisted retrieval workflow opened:

- `tool-metadata-verified`: authoritative title/author/venue metadata was opened;
- `tool-abstract-verified`: an abstract was also opened;
- `tool-full-text-verified`: the full relevant source text was opened;
- `tool-description-verified`: authoritative publisher description was opened for a book or framework.

The quoted evidence is a verbatim title, abstract, significance, or description line, identified by the row. These labels do **not** mean a human author completed the final source check. Human re-verification of all load-bearing citations remains mandatory before submission.

Ref.	Verification state	Review status / journal score	Opened evidence and relevance
[1]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed journal, **4/5**	"A Self-Organizing Spatial Vocabulary." Establishes local vocabulary self-organization.
[2]	**tool-metadata-verified**, arXiv title and DOI metadata	Peer-reviewed journal, **4/5**	"Sharp transition towards shared vocabularies in multi-agent systems." Supports convergence dynamics.
[3]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed NeurIPS paper; journal score N/A	"Learning to Communicate with Deep Multi-Agent Reinforcement Learning." Establishes RIAL/DIAL.
[4]	**tool-metadata-verified**, arXiv primary page	Peer-reviewed NeurIPS paper; journal score N/A	"Learning Multiagent Communication with Backpropagation." Establishes continuous learned communication.
[5]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed ICLR paper; journal score N/A	"Multi-Agent Cooperation and the Emergence of (Natural) Language." Establishes referential-game protocol learning.
[6]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed NeurIPS paper; journal score N/A	"Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols." Supports sequence messages.
[7]	**tool-metadata-verified**, ACL Anthology	Peer-reviewed EMNLP paper; journal score N/A	"Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog." Supports the degenerate-code caution.
[8]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed AAAI paper; journal score N/A	"Emergence of Grounded Compositional Language in Multi-Agent Populations." Supports grounded partial composition.
[9]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed ICLR paper; journal score N/A	"Compositional Obverter Communication Learning From Raw Visual Input." Supports raw-input grounding with a designed bias.
[10]	**tool-metadata-verified**, ACL Anthology	Peer-reviewed EMNLP paper; journal score N/A	"How agents see things: On visual representations in an emergent language game." Supports success-versus-grounding caution.
[11]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed AAMAS paper; journal score N/A	"On the Pitfalls of Measuring Emergent Communication." Supports positive signaling/listening distinction.
[12]	**tool-metadata-verified**, arXiv primary page	Peer-reviewed NeurIPS paper; journal score N/A	"Biases for Emergent Communication in Multi-agent Reinforcement Learning." Shows designed biases can promote communication.

Ref.	Verification state	Review status / journal score	Opened evidence and relevance
[13]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed AAMAS paper; journal score N/A	"Capacity, Bandwidth, and Compositionality in Emergent Language Learning." Supports capacity/bandwidth controls.
[14]	**tool-metadata-verified**, ACL Anthology	Peer-reviewed ACL paper; journal score N/A	"Compositionality and Generalization In Emergent Languages." Supports separate behavioral generalization tests.
[15]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed NeurIPS paper; journal score N/A	"Interpretable agent communication from scratch (with a generic visual processor emerging on the side)." Supports causal symbol interventions.
[16]	**tool-full-text-verified**, PMC full text and Crossref	Peer-reviewed PNAS journal article, **5/5**	"Words categorize the semantic fields they refer to in ways that maximize communication accuracy while minimizing complexity." Supports discrete-channel efficiency in the tested color domain.
[17]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed ICLR paper; journal score N/A	"Compositional Languages Emerge in a Neural Iterated Learning Model." Supports transmission bottlenecks.
[18]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed IEEE journal article, **4/5**	"Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity." Establishes computational iterated learning.
[19]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed PNAS journal article, **5/5**	"Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language." Establishes human laboratory transmission effects.
[20]	**tool-abstract-verified**, PMLR primary page	Peer-reviewed ICML paper; journal score N/A	"Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning." Supports causal social-influence reward.
[21]	**tool-abstract-verified**, PMLR primary page	Peer-reviewed ICML paper; journal score N/A	"Curiosity-driven Exploration by Self-supervised Prediction." Supports curiosity as intrinsic reward.
[22]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed NeurIPS paper; journal score N/A	"Learning to Draw: Emergent Communication through Sketching." Supports graphical-carrier feasibility.
[23]	**tool-metadata-verified**, ACL Anthology	Peer-reviewed EMNLP paper; journal score N/A	"Emergent Linguistic Phenomena in Multi-Agent Communication Games." Supports community contact and protocol convergence.
[24]	**tool-abstract-verified**, arXiv primary page/API and venue comment	Peer-reviewed NeurIPS paper; journal score N/A	"Emergent Communication: Generalization and Overfitting in Lewis Games." Used as supporting, not sole, evidence.
[25]	**tool-metadata-verified**, arXiv primary page	Peer-reviewed ICLR paper; journal score N/A	"On the role of population heterogeneity in emergent communication." Supports heterogeneity effects.
[26]	**tool-metadata-verified**, official journal page	Peer-reviewed journal article, **3/5**	"Learning and communication pressures in neural networks: Lessons from emergent communication." A field review with a newer/variable-impact venue.
[27]	**tool-metadata-verified**, ACL Anthology	Peer-reviewed COLING paper; journal score N/A	"Searching for Structure: Investigating Emergent Communication with Large Language Models." Direct pretrained-LLM comparison.
[28]	**tool-abstract-verified**, arXiv page and TMLR record metadata	Peer-reviewed TMLR journal article, **4/5**	"A Review of the Applications of Deep Learning-Based Emergent Communication." Used as a field synthesis, not primary experiment.
[29]	**tool-abstract-verified**, arXiv primary page/API	Peer-reviewed ICLR paper; journal score N/A	"Emergent Communication through Negotiation." Supports incentive-sensitive communication.
[30]	**tool-metadata-verified**, Crossref/JSTOR metadata; full page blocked	Peer-reviewed Econometrica journal article, **5/5**	"Strategic Information Transmission." Supplies cheap-talk theory.
[31]	**tool-metadata-verified**, Crossref DOI metadata; publisher page blocked	Peer-reviewed Science journal article, **5/5**	"Human-level play in the game of Diplomacy by combining language models with strategic reasoning." Strategic-communication precedent only.
[32]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed Artificial Life journal article, **4/5**	"The Development of Embodied Cognition: Six Lessons from Babies." Supports embodied-development framing.
[33]	**tool-abstract-verified**, Duke publication page and Crossref	Peer-reviewed Child Development journal article, **5/5**	"Joint attention and early language." Supports the relevance of joint attention in human development.

Ref.	Verification state	Review status / journal score	Opened evidence and relevance
[34]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed IEEE journal article, **4/5**	"Intrinsic Motivation Systems for Autonomous Mental Development." Supports learning-progress mechanisms.
[35]	**tool-description-verified**, Crossref and Google Books	Scholarly press book; journal score N/A	"Raising AI." Publisher description: "a framework of empowerment for a future with our artificial children." Used as normative metaphor, not empirical evidence.
[36]	**tool-abstract-verified**, University of Bath publication record	Peer-reviewed journal article, **4/5**	"Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." Supports caution about assigning moral status through metaphor.
[37]	**tool-metadata-verified**, AAAI/ACM publication metadata	Peer-reviewed conference paper; journal score N/A	"All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI." Supports anthropomorphism-risk controls.
[38]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed PNAS journal article, **5/5**	"The preregistration revolution." Supports prediction/postdiction separation.
[39]	**tool-abstract-verified**, University of Bristol record and Crossref	Peer-reviewed Nature Human Behaviour review, **5/5**	"A manifesto for reproducible science." Supports methods, reporting, and incentive reforms.
[40]	**tool-abstract-verified**, JMLR abstract page	Peer-reviewed journal article, **5/5**	"Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." Supports code, checklist, and challenge practices.
[41]	**tool-metadata-verified**, Crossref DOI metadata; publisher page blocked	Peer-reviewed Journal of Cryptology article, **4/5**	"How to time-stamp a digital document." Supports private tamper-evident timestamping.
[42]	**tool-full-text-verified**, RFC Editor	IETF Experimental RFC; journal score N/A	"Certificate Transparency." Defines publicly auditable append-only Merkle logs.
[43]	**tool-full-text-verified**, RFC Editor	Informational RFC; journal score N/A	"JSON Canonicalization Scheme (JCS)." Defines invariant JSON for repeatable cryptographic operations.
[44]	**tool-description-verified**, NIST official page/PDF	Government consensus framework; journal score N/A	"Artificial Intelligence Risk Management Framework." Supports lifecycle risk management.
[45]	**tool-metadata-verified**, Crossref DOI metadata	Peer-reviewed Nature Machine Intelligence article, **5/5**	"The global landscape of AI ethics guidelines." Supports convergence/implementation-gap framing.
[46]	**tool-full-text-verified**, CoLab publication page	Institutional report, not peer-reviewed; **1/5** under the fixed journal rubric	"What Is Ethical AI?" Used for CoLab governance commitments, not independent empirical evidence.
[47]	**tool-full-text-verified**, CoLab publication and source repository	Institutional research guidance, not peer-reviewed; **1/5**	"AI-Powered Assistance in Formulating Research Questions." Supplies the source-verification and AI-disclosure method.
[48]	**tool-metadata-verified**, arXiv primary page	Preprint, not peer-reviewed; journal score N/A	"Learning to Protect Communications with Adversarial Neural Cryptography." Motivates an exploratory synthetic-message experiment only.
[49]	**tool-metadata-verified**, ACL Anthology	Peer-reviewed EMNLP demo paper; journal score N/A	"EGG: a toolkit for research on Emergence of lanGuage in Games." Closest infrastructure comparator.
[50]	**tool-abstract-verified**, PMLR primary page	Peer-reviewed ICML paper; journal score N/A	""Other-Play" for Zero-Shot Coordination." Supports novel-partner evaluation and specialized-convention risk.

No source in the table was silently upgraded from preprint or workshop status to a peer-reviewed main-track result.

Appendix B. Ethical Tech CoLab Journal Credibility Rubric

This is the fixed rubric used by the CoLab's Researcher Prompt [47]. The economics examples are illustrative; the score-band definitions are unchanged.

Score	Criteria	Interpretation in this manuscript
5/5	Top-tier, highly selective journals with rigorous peer review and high impact; widely recognized in the field.	Examples here include Science, PNAS, Nature Human Behaviour, Nature Machine Intelligence, JMLR, Econometrica, and Child Development.
4/5	Well-regarded field-specific journals with strong peer review and consistent citation impact.	Examples here include Artificial Life, IEEE Transactions on Evolutionary Computation, Ethics and Information Technology, Journal of Cryptology, JSTAT, and TMLR.
3/5	Reputable peer-reviewed journals with variable impact or lower selectivity, often open access.	Used here for the newer field journal Language Development Research.
2/5	Limited peer review or editorial oversight; may include some proceedings or trade publications.	No journal source is assigned this band in the current corpus.
1/5	Non-peer-reviewed sources, blogs, or promotional/institutional materials; unsuitable as sole academic evidence.	CoLab institutional guidance is scored here and used only for declared house rules and governance framing.

Premier peer-reviewed computer-science conference papers are labeled as such but are not assigned a journal score because the fixed rubric is explicitly journal-based.

Appendix C. Required Work Before arXiv Submission

- [ ] Finalize title, author order, affiliations, ORCIDs, and corresponding author.
- [ ] Obtain or document ethics/governance review.
- [ ] Restore Tavily access and rerun the literature search.
- [ ] Run a systematic scholarly-database search with recorded query strings, inclusion/exclusion criteria, and dual screening.
- [ ] Confirm peer-review status and final bibliographic metadata for every source.
- [ ] Pre-register primary hypotheses, sample sizes, seeds, and analysis code.
- [ ] Complete E00-E03 qualification before substantive experiments.
- [ ] Complete the experiment-specific integrity and verification checks.
- [ ] Replace blank results tables only with verified outputs.
- [ ] Report all exclusions, invalid runs, deviations, null findings, and negative results.
- [ ] Conduct an independent statistical review.
- [ ] Conduct an independent citation and claim audit.
- [ ] Add figures generated from verified evidence, with code and data references.
- [ ] Freeze the manuscript, code, data, and evidence-bundle commit hashes.
- [ ] Add funding, conflicts, author contributions, and acknowledgements.
- [ ] Decide the text license and confirm permission for every reproduced figure.
- [ ] Generate the arXiv TeX/PDF package and validate accessibility.
- [ ] Ensure the abstract and conclusion describe results no more strongly than the evidence supports.

Appendix D. Worked Preregistration Example: E03 Controls

> **Status:** Complete worked example for methodological review; not yet registered or executed.

D.1 Registration and Integrity

- Registration target: OSF Registries.
- The canonical registration JSON, this manuscript commit, analysis-script commit, and generated seed manifest will be registered before execution.
- The same `preRegistrationHash` will be anchored to Base Sepolia before any run enters `running`.
- No outcome will be inspected before registration and anchoring complete.

D.2 Objective

Test whether a four-choice referential task can be solved above chance without meaningful learned communication and establish an oracle upper bound.

D.3 Fixed Configuration

Parameter	Registered value
Experiment	`E03-v1`
Deployment mode	Research-Grade Mode
Learner track	`no-learning`
Interaction	Cooperative signaling
Carrier	Fixed token
Symbol inventory	32 symbols
Maximum message length	4 symbols
Affect	Disabled
Candidate objects	4 per episode
Chance success	0.25
Episodes per run	200
Primary seeds per condition	Power-rule output; 75 when pilot SD is greater than 0.05 and at most 0.10
Reserve seeds per condition	10% of primary count, rounded up
Conditions	Disabled, constant, random, shuffled, normal no-learning, oracle
Alpha	0.05
Within-family correction	Holm-Bonferroni
Equivalence margin	+/-0.05 success probability

D.4 Seed Generation

For slot `i`, the shared scenario seed is the hexadecimal SHA-256 digest of:

```
ald-e03-v1 || 0x00 || decimal(i)
```


Slots `1` through registered `N` are primary. The next `ceil(0.10 * N)` slots are ordered reserves. Every condition uses the same scenario seed for a given slot. Condition-specific gateway randomness for `random` and `shuffled` is derived from:

```
scenarioSeed || 0x00 || communicationCondition
```

The complete scenario and gateway seed manifest is part of the registration.

D.5 Conditions

1. **Disabled:** no artifact is delivered.
2. **Constant:** the same pre-registered valid artifact is delivered every episode.
3. **Random:** a seeded random valid artifact is delivered.
4. **Shuffled:** an artifact from another episode in the same registered evaluation batch is delivered according to a seeded permutation.
5. **Normal no-learning:** the fixed no-learning policy's valid proposal is delivered.
6. **Oracle:** the deterministic Scenario Engine delivers the minimal sufficient artifact from researcher-only ground truth.

The channel event records proposal and delivered-artifact hashes where a learner proposal exists.

D.6 Primary Outcomes and Tests

The unit of analysis is the run/seed success proportion across 200 episodes.

1. **Control equivalence:** For each of disabled, constant, random, shuffled, and normal no-learning, perform two one-sided one-sample tests on seed-level success proportions against equivalence bounds 0.20 and 0.30. Equivalence requires both one-sided tests to reject at the Holm-adjusted alpha.
2. **Oracle adequacy:** The lower bound of the two-sided 95% bootstrap confidence interval for mean seed-level oracle success must exceed 0.90.
3. **Oracle separation:** For each non-oracle condition, compute paired seed-level oracle-minus-control differences. The lower bound of the Holm-adjusted 95% confidence interval must exceed 0.60.

No episode is analyzed as an independent run.

D.7 Sensitivity and Power

Before final registration, a separate outcome-blind-for-confirmatory-use pilot of 20 seeds per non-oracle condition will estimate the largest between-seed standard deviation. Pilot runs will not enter confirmatory estimates. The registered primary seed count is selected by this fixed rule:

Largest pilot SD	Primary seeds per condition
<= 0.05	25
> 0.05 and <= 0.10	75

Largest pilot SD	Primary seeds per condition
> 0.10 and <= 0.15	150
> 0.15 and <= 0.20	300
> 0.20	New simulation and amended registration required before collection

A 30,000-replicate Monte Carlo design check was run for this draft under a true seed-level mean of 0.25, between-seed standard deviation of 0.10, 200 binomial episodes per seed, and conservative per-test alpha of 0.01. Estimated equivalence-test power was 0.924 at 75 seeds per condition. Sensitivity checks produced approximately 0.912 power at SD 0.05 with 25 seeds, 0.920 at SD 0.15 with 150 seeds, and 0.913 at SD 0.20 with 300 seeds. Before registration, the simulation code and output must be checked in and independently rerun. Failure to reproduce at least 90% power blocks registration; it does not permit post-hoc widening of the margin.

Sensitivity analyses:

- Wilson intervals over pooled episodes are descriptive only;
- a hierarchical Bernoulli model with seed random intercept is reported as a robustness check;
- invalid primary runs are treated as failures in a worst-case sensitivity analysis.

D.8 Exclusions, Invalid Runs, and Replacement

A run is invalid only for a pre-specified verifier failure, wrong configuration hash, gateway-control mismatch, or infrastructure failure that prevents the planned 200 episodes.

- Invalid runs remain in the public run index.
- The next unused reserve seed for that condition replaces the invalid primary slot.
- No more than the registered reserve count may be used per condition.
- If fewer than registered `N` valid runs remain, the primary analysis is incomplete; there is no unregistered additional collection.
- Invalid and aborted counts are reported by condition.

D.9 Stopping Rule

All registered primary slots are attempted. There is no efficacy-based early stopping, optional extension, or outcome-dependent seed replacement.

D.10 Decision Rule

E03 qualifies the downstream chance baseline only if:

- all five non-oracle conditions meet equivalence;
- oracle adequacy and separation criteria pass;
- no non-oracle condition has more than 5% of its primary seeds with observed success of 0.35 or greater; every such seed is individually audited for leakage;
- all included evidence bundles pass verification;
- no unplanned metadata or channel leakage is detected.

Failure is a result and blocks downstream confirmatory claims until a new, separately registered protocol is justified.

Appendix E. Future Results-Table Scaffolds

These tables contain no observed results.

E.1 Qualification

Experiment	Planned runs	Invalid runs	Valid runs	Primary criterion	Status	Evidence root
E00 Integrity	TBD	—	—	Detect all mutation classes	Not run	—
E01 Isolation	TBD	—	—	Block all enumerated side routes	Not run	—
E02 Leakage	TBD	—	—	Rule out pre-registered leakage bound	Not run	—
E03 Controls	6 x registered N (450 at current SD assumption)	—	—	Equivalence and oracle bounds	Not run	—

E.2 Core Emergence

Condition	Seeds	Invalid runs	Held-out success	Positive listening	Ledger agreement	Generalization
No learning	—	—	—	—	—	—
Frozen LLM	—	—	—	—	—	—
Scratch RL, extrinsic	—	—	—	—	—	—
Scratch RL, intrinsic	—	—	—	—	—	—
Self-supervised	—	—	—	—	—	—
Hybrid	—	—	—	—	—	—

E.3 Carrier and Affect

Condition	Seeds	Invalid runs	Convergence time	Stable forms	Repair success	Leakage	Held-out success
Fixed tokens	—	—	—	—	—	—	—
Unfamiliar glyphs	—	—	—	—	—	—	—
Bitmap	—	—	—	—	—	—	—
Vector strokes	—	—	—	—	—	—	—
Tones	—	—	—	—	—	—	—
Six-display affect	—	—	—	—	—	—	—

E.4 Transfer and Negotiation

Condition	Seeds	Invalid runs	Zero-shot transfer	Adapted transfer	Informativeness	Agreement	Ambiguity
Original partner	—	—	—	—	—	—	—
New seed, same architecture	—	—	—	—	—	—	—
New architecture	—	—	—	—	—	—	—

Condition	Seeds	Invalid runs	Zero-shot transfer	Adapted transfer	Informativeness	Agreement	Ambiguity
Cooperative signaling	—	—	—	—	—	—	—
Semi-cooperative negotiation	—	—	—	—	—	—	—
Conflicting negotiation	—	—	—	—	—	—	—
No-agreement control	—	—	—	—	—	—	—