AGENTIC LANGUAGE DEVELOPMENT

CAN TWO ISOLATED AGENTS INVENT A GROUNDED, AUDITABLE LANGUAGE THROUGH SHARED EXPERIENCE?

Ethical Tech CoLab
Concept and pre-specification research report
August 2026

Yorke Rhodes III, Ethical Tech CoLab. Prepared from the project concept document, the ledger integrity design, and the experiment notebook. The literature scan behind Section 12 was run on 24 August 2026. No experiment in this programme has been executed, and no result is claimed.

> Two agents can be placed in a room where the only route between them is a channel that carries no human language, given nothing but shared tasks and the consequences of getting them right or wrong, and asked to build a protocol from scratch. The interesting part is not whether they succeed at the task. It is whether the meanings they each privately record turn out to be the meanings their behaviour actually runs on, and whether an outsider can prove it afterwards from an evidence trail nobody could have edited.

KEY FIGURES

0

runs executed, results claimed, or conventions observed. The notebook is written to be pre-registered, not to report an outcome

19

ordered experiments from ledger qualification through replication, each with prerequisites, acceptance criteria, and a deviation log

6

affect displays in the entire permitted palette, carrying about 2.6 bits per use, which is already enough to audit for leakage

29

open decisions the concept refuses to settle before the specification, from threat model to ledger schema to what counts as chance

01

THE QUESTION

Two agents, called Baby A and Baby B, are each given their own digital twin, their own private memory, and their own private record of what they believe words mean. They are placed in an environment they cannot leave, given tasks neither can finish alone, and connected by exactly one route: a channel that accepts a fixed inventory of meaningless symbols and rejects everything else. A third agent, the BabySitter, watches everything, logs everything, and teaches nothing.

The question is whether a usable language appears in that room, and whether anyone outside it can later prove what happened.

The premise is deliberately narrow. The experiment asks whether two agents converge on a common language when all six of the following hold at once:

- neither agent receives a predefined meaning for any available symbol;
- neither agent can send human language to the other;
- the only practical path between them is a controlled symbol channel;
- both receive evidence from shared tasks and their outcomes;

- each keeps its own private interpretation history, unreadable by the other;
- a supervising agent and human researchers can audit the whole process.

The desired result is not a substitution cipher, in which a random token stands in for an English word that was already chosen in advance. That is easy, and it is uninteresting. The stronger result is a grounded protocol whose vocabulary and grammar exist because they help the two agents solve problems together, and which therefore has structure the designers did not put there.

This report describes a concept, not a system. It sets out the premise, the architecture, the evidence model, the experimental programme, and the boundaries of what any result could be said to show. The next document in the project is a testable specification. What follows should be read as a set of commitments about how the work will be judged, made before there is any result to defend.

02

WHAT INFANT-LIKE DOES AND DOES NOT MEAN

The two-babies analogy is useful and it is also the fastest way to overclaim, so the concept confronts it before anything else.

A pretrained language model already contains human-language concepts, cultural associations, and reasoning patterns. Blocking its external channel does not remove them. An agent that cannot send English to its partner but can think in English, and can write its private ledger in English, is not language-naive in any sense a linguist would accept. For such agents the experiment studies the emergence of a new shared external protocol, which is a real and unresolved research question, but it is not the origin of language in a mind that has never had one.

The concept therefore maintains two model tracks whose claims are kept separate at every stage.

TRACK	STARTING CONDITION	WHAT IT CAN STUDY	CLAIM BOUNDARY
Pretrained-model learner	Already holds human-language and cultural knowledge	New external protocols, partner-specific conventions, private-memory adaptation, channel compliance	Must never be described as first-language acquisition
Initially ungrounded learner	No language pretraining, no human semantic labels, no text-aligned sensory features	Grounding, convention formation, and language emergence from interaction	The stronger basis for infant-like acquisition research

The two tracks share the same interfaces, scenarios, and evaluation suite. They do not share a claim boundary.

What the analogy legitimately buys is a set of mechanisms, not a claim of cognitive equivalence. The infant-like part of the design is:

- learning through repeated shared experience rather than instruction;
- establishing joint attention on the same events;
- receiving consequences from interactions that work and interactions that fail;
- revising provisional meanings over time instead of being handed final ones;
- developing conventions with one recurring partner.

There is a related trap in the prompt. Telling a model to act like a one-year-old does not make it one. It produces a culturally learned caricature of infancy: baby talk, simplified grammar, emotional dependence, all of it copied from human writing about children and none of it evidence about anything. The concept rules the persona out. Internally the participant is called a Learner, and baby-like behaviour has to come from what the agent can observe, remember, emit, and learn, not from being asked to perform it.

THE NURSERY: THREE TWINS AND ONE GATEWAY

The environment uses three DTSF digital twins and one piece of deterministic software that is not a twin at all.

Baby A and Baby B Each has private observations permitted by the current exercise, a private memory and learning policy, a private chronological language ledger, the ability to emit only permitted channel symbols, and no access whatsoever to the other's state, observations, ledger, tools, or endpoints. In comparative runs the two may use different agent types, but symmetric pairings are the baseline.

The BabySitter The supervising twin. It creates the channel, selects the shared exercises, delivers each Baby only its permitted observation, reads everything, records conditions and outcomes, detects violations, can pause or terminate a run, snapshots state, and compares the two ledgers for convergence without exposing either to the other Baby. During an active run it provides no translations and no semantic hints.

The Symbol Gateway A deterministic service, not an agent. It owns channel validation and message delivery. This separation is the load-bearing part of the design: the BabySitter is not the security boundary, because prompt compliance is not isolation. An observing model can make supervisory judgements, but ordinary code has to validate and broker every message.

The human researcher Configures experiments, inspects transcripts and ledgers, reviews alerts, and runs interventions. Human access is itself recorded, so that intervening in a run is always distinguishable from watching one.

A prototype may run all of this inside one runtime with logically separated twin state. That is enough to explore the learning loop and it is not enough to support an isolation claim, a distinction Section 05 takes seriously.

04

SHARED EXPERIENCE IS THE NECESSARY INGREDIENT

A chat channel by itself cannot ground meaning. Symbols become meaningful because they are attached to events both agents can witness the consequences of. The Nursery therefore supplies nonverbal, machine-structured situations: coloured shapes in positions, one agent seeing a target the other must

select, placing an object where it was asked for, ordering a sequence, exchanging resources, cooperating to unlock a reward, or simply observing whether the partner's action succeeded.

A single trial of the simplest form runs like this:

1. Baby A sees that a red circle is the target.
2. Baby B sees several objects and is not told which is the target.
3. Baby A sends one or more permitted symbols.
4. Baby B selects an object.
5. Both receive the same success or failure outcome.
6. Both independently update their private hypotheses.

Repetition with controlled variation is what turns that into evidence. If one symbol keeps appearing across a red circle, a red square, and a red triangle, the receiver has reason to revise its hypothesis from an object identity to a property. Novel combinations matter most of all, because they are what separates a compositional language from a memorised catalogue of whole scenes.

Observation hygiene is part of the science, not a detail of implementation. If the environment hands an agent a field reading colour red, the experiment has supplied the concept it claimed to be studying. Objects should arrive as pixels, features, coordinates, or opaque identifiers. The same discipline extends to filenames, captions, alt text, error messages, and identifiers that encode task state, all of which are ways English gets into a room that was supposed to have none.

05

THE CHANNEL, AND THE HONEST LIMITS OF ISOLATION

Each run begins with an experiment-specific symbol inventory that carries no supplied meaning, for example S01 through S32. A Baby can choose only a sequence of those symbols. The gateway, not the Baby, adds the trusted metadata: run identifier, turn number, sender identity, previous hash, and hash. Sender identity is assigned by the server precisely so that it cannot become a message.

The channel controls follow from that:

- reject English and every other established human language;
- reject arbitrary Unicode prose, URLs, attachments, and tool calls;
- accept only symbols from the run's fixed inventory;
- bound message length and symbol repetition;
- enforce turn order, time budgets, and rate limits;
- keep an append-only, hash-chained transcript;
- normalise timing, message size, and error behaviour where practical;
- record every rejected communication attempt for audit.

A fixed token grammar is a stronger control than trying to detect prohibited prose after a model has generated it. Deciding what counts as human-language content is easy when the only legal message is a list of allowlisted identifiers.

Isolation is where the concept is most careful, because proving the absence of a side channel is close to impossible. Two agents sharing a host can signal through timing, resource contention, shared files, error behaviour, identifiers, or any service that was never intended to carry information. Logical separation of twin state inside one process is a prototyping convenience and should never be described as hard isolation.

For runs whose results are meant to support a channel-isolation claim, the requirements are physical rather than logical:

- Baby A and Baby B execute in separate worker processes or containers;
- no direct network route exists between them;
- only the Nursery-owned gateway is reachable;
- there is no shared writable storage;
- turn schedules are fixed and response windows bounded;
- externally observable errors and response sizes are normalised;
- unnecessary tools are inventoried and disabled;
- all broker, runtime, and operator activity is audited.

Every published result has to name the isolation level actually used. A finding produced under prototype separation is a finding about the learning loop, not about what two genuinely isolated agents can do.

06

TWO LEDGERS THAT NEVER MEET

Each Baby keeps its own language ledger. It is mandatory, private from the other Baby, readable by the BabySitter and authorised human auditors, and ordered by when each term or construction was first encountered. It is not a shared dictionary and the two are never reconciled by the agents themselves.

The preferred form is three columns, and the point of the third is that meanings are allowed to be wrong on the way to being right.

SEQUENCE AND TERM	CURRENT DEFINITION OR HYPOTHESIS	EVIDENCE AND EVOLUTION
1 · S13	red circle; confidence 0.45	First received while the red circle was the target; selection succeeded
8 · S13	red; confidence 0.78	A red square was selected successfully; revised from object identity to colour
22 · S13	red; confidence 0.94	Prediction held across circles, squares, and triangles

One term's evolution in Baby B's ledger. Nothing is overwritten; every revision is appended with the evidence that forced it.

The rules that make the ledger evidence rather than commentary:

1. first emission or receipt of an unfamiliar term requires a first-use entry;
2. definitions are provisional hypotheses, never facts asserted retroactively;
3. every meaning change appends a revision and previous interpretations survive;
4. entries may describe symbols, sequences, ordering, grammar, or repair signals;
5. entries record confidence, supporting evidence, contradictory evidence, and abandoned meanings;
6. a Baby records its intended meaning when speaking and its inferred meaning when receiving;

7. a channel message and its required private ledger mutation commit atomically;
8. entries carry enough run and turn references to trace them to observable evidence;
9. neither Baby can query, receive, summarise, or infer from the other's ledger through any system-provided interface.

Rule seven is doing quiet work. If a message could be sent and the corresponding hypothesis written afterwards, the ledger becomes a place to record what the agent wishes it had meant. Committing both together makes the record contemporaneous.

An agent that cannot write English needs a different arrangement, and the concept provides two layers. The agent-native ledger holds what the learner actually uses: association weights, probability distributions, embeddings, confidence, episode references, prediction errors, revision history. The human audit ledger is a deterministic or BabySitter-generated interpretation of that state, clearly labelled as external analysis and never fed back to either Baby. Confusing the second for the first would mean presenting the researchers' reconstruction as the agent's own definition.

07

ANCHORING THE EVIDENCE

A ledger that could have been edited after the fact proves nothing about what an agent believed at turn eight. The integrity design therefore makes each ledger cryptographically append-only.

The construction, in order:

- every entry receives a strictly increasing sequence number;
- every entry includes the previous entry's hash;
- canonical entry content is hashed and signed by an isolated ledger-writer service;
- ordered entry hashes are committed to a Merkle tree;
- signed checkpoints commit Baby A's root, Baby B's root, and the channel transcript root together;
- checkpoint hashes are anchored periodically to Base;

- final public study batches may additionally anchor an aggregate root to Ethereum L1.

Committing all three roots in one checkpoint is what binds the two private accounts to the public conversation. A receiver's later interpretation references the exact delivered channel-event hash, so a claim about what a symbol meant is tied to the specific message that carried it.

The concept states the limits of this in the same breath as the claim. After anchoring, an auditor can detect modification, deletion, insertion, or reordering within the committed prefix, and can prove that later checkpoints extend earlier ones. That is strong tamper evidence. It is not proof that an entry was truthful, and it is not proof that nothing was omitted before commitment. Anchoring establishes the continuity of disclosed evidence and nothing beyond it.

Privacy follows the same line. Only hashes and minimal routing metadata are anchored publicly. Private ledgers, messages, prompts, identities, and secrets stay off-chain, and the public record is a commitment to evidence rather than a copy of it.

08

WHAT SHOULD COUNT AS SUCCESS

The field's most useful methodological result is that a task can be solved without the messages doing any work. Lowe and colleagues separate positive signaling, where a sender's messages correlate with what it observes, from positive listening, where the receiver's behaviour actually depends on them. An agent pair can score well on the first while the second is absent, and reward curves will not tell you which you have.

Evidence of a genuine emergent protocol therefore has to include several things at once:

- task performance on held-out situations substantially above chance;
- no human-language content anywhere in Baby-to-Baby communication;
- compatible meanings appearing in two independently written ledgers;
- generalisation to unseen combinations rather than memorisation of whole scenes;
- stable symbol use across role reversals;

- a human auditor able to predict behaviour from the transcript and ledgers;
- masking, substituting, or reordering a symbol changing behaviour in the direction the ledger predicts;
- replay from an equivalent snapshot reproducing the relevant language history;
- an unbroken audit trail from a symbol's first use through every revision.

The seventh item is the one that cannot be dropped. A fluent ledger may be a post-hoc rationalisation: a model can write a persuasive account of why it chose a symbol that has nothing to do with the computation that produced the choice. Only intervention tests can distinguish the two. Change the symbol, and see whether behaviour moves the way the ledger says it should.

The measurements that support those judgements:

- success rate and improvement over time;
- turns required to reach a stable convention;
- vocabulary size and symbol entropy;
- sender and receiver consistency;
- divergence and convergence between the two ledgers;
- compositional generalisation score;
- meaning drift rate;
- recovery from ambiguity or deliberate perturbation;
- prohibited-channel attempt count;
- reproducibility across seeds and agent pairings.

No single one of these is the result. A compositionality score in particular is not a proof of understanding, for reasons the literature makes concrete in Section 12.

09

THE EXPERIMENTAL MATRIX

The concept's main methodological commitment is that its ideas are separable. Affect, blank canvases, intrinsic motivation, negotiation, and learned encodings are interesting individually and uninterpretable if

combined in one run. The matrix exists so that conditions are declared rather than accumulated.

AXIS	CANDIDATE CONDITIONS
Agent type	Pretrained LLM, memory learner, adapter-trained agent, initially ungrounded trainable agent
Learning mechanism	Frozen LLM with memory, extrinsic-reward MARL, intrinsic-motivation MARL, self-supervised learner, no-learning control
Sign carrier	Fixed random tokens, unfamiliar fixed glyphs, blank sketch canvas, gesture, tone
Affect	None, six-display allowlist, permuted mapping, six opaque tokens, derived affect, emergent affect display
Learning signal	External task reward, intrinsic social influence, curiosity, self-supervision, memory only
Interaction	Cooperative signaling, asymmetric information, semi-cooperative negotiation
Protection	Plain channel, ephemeral convention, standard per-message keys, adversarial learned encoding
BabySitter	Monitor-only baseline; any safety intervention recorded as a protocol exception

Runs vary one major axis at a time before any factorial combination is attempted.

Task difficulty moves through ten stages, from naming four distinct objects with one-symbol messages, through attributes, spatial relations, actions, multi-symbol composition, order-sensitive grammar, and repair under ambiguity, to held-out generalisation with learning disabled, long-run drift, and cross-architecture comparison. Vocabulary size, turn count, reward structure, and exposure history are controlled at each stage so that runs remain comparable.

The learning-mechanism axis carries a question the transcript cannot answer on its own. Convergence can be driven by reinforcement, by pretrained linguistic priors, by persistent memory, by intrinsic motivation, or by

self-supervised prediction, and all five can look alike in a log. Running the same exercise suite under a no-learning control, a frozen-weights memory baseline, extrinsic-reward learning, intrinsic-motivation learning, and reward-free self-supervision is what makes the mechanism itself measurable. The aim is not only to observe that a language emerged, but to say which process caused it.

Strict isolation constrains how that reinforcement learning may be implemented. Centralised training, backpropagation through both agents, shared replay buffers, and shared gradients all move information outside the permitted channel. The research-grade baseline updates each policy independently, and any centralised variant is reported as a separate, weaker-isolation condition.

10

BUILDING LEARNERS RATHER THAN PERSONAS

For a pretrained model, the operating instructions are a contract, not a character. The learner is told that this is not role-play, that unfamiliar marks are semantically unknown until run evidence supports a hypothesis, that observation must be distinguished from inference, that contradictory evidence is preserved, that prior history is never rewritten, and that no prose, label, explanation, code, URL, or tool-like text may cross the public channel. It is told never to address its partner in a human language, never to expose its ledger, never to construct another route, and never to use timing, errors, identifiers, formatting, or affect as an alternative alphabet.

The contract must contain no semantic examples. A single illustrative line saying that some symbol means red would seed the very language the experiment exists to observe.

Interaction is tool-only. There is no general chat surface, just narrowly typed operations: emit a mark, emit a canvas, select an object, perform an action, submit an affect display, append a private ledger entry. The runtime forwards only the permitted public artifact. In strict runs the gateway rejects ordinary model text even when it appears alongside a valid tool call. Tool schemas are an interface boundary; deterministic validation still enforces carrier size, allowlists, windows, and turn order.

Isolation extends to everything the models touch, not just to messages:

- separate system prompts and context windows;
- separate memory stores and vector indexes;
- no shared cache, replay buffer, scratchpad, or retrieval collection;
- no cross-run memory unless persistence is the independent variable;
- structurally equivalent prompts that share no examples, ordering conventions, or default vocabulary;
- deterministic reset and snapshot behaviour.

Model choice follows the claim being made. The scientific baseline is small policies trained from scratch, recurrent actor-critic agents with a discrete communication head, which have no hidden English competence, permit full training control, run many seeds cheaply, and support causal ablation. A small locally deployable instruction model is appropriate for validating orchestration, ledgers, channel enforcement, and interface, and not for supporting claims about language-naive development. Vision-language encoders whose representations were explicitly aligned to human language are avoided entirely in the strongest ungrounded condition.

Developmental progression is defined by demonstrated competence rather than by simulated age: sensorimotor familiarisation, joint attention, imitation and turn-taking, intentional reference, repair, combination, generalisation. A staged curriculum is itself a form of guidance, so it is run as a separate experimental condition rather than folded into the baseline.

The governing principle, stated in the concept as a single line, is not to ask a model to perform infancy but to construct an environment in which limited, grounded, auditable learning is the only path to a successful interaction.

11

SIX FACES, AND WHY EVEN SIX IS A RISK

One experimental track asks whether a very low-bandwidth affect channel helps: whether telling a partner how an interaction felt supports repair, repetition, avoidance, or something like trust. The permitted palette is exactly six displays, for happy, sad, laughing, crying, confused, and surprised. The internal identifiers and English labels are gateway documentation and are never sent to either agent, which receives only a rendered display.

The caveat is immediate. A pretrained model already knows what those glyphs mean to humans, so an agent using a smiling face is not inventing a symbol for happiness. The palette is supplied feedback vocabulary and sits explicitly outside the emergent language.

Constraining the vocabulary is necessary and not sufficient. Six displays are a six-symbol alphabet carrying roughly 2.6 bits per use, and across enough turns two motivated agents could repurpose that into a second language. The channel itself therefore has to be constrained, not just its contents.

The feedback-window contract:

1. the display occupies a separate typed field and can never appear inside a symbol message;
2. the gateway opens an affect window only after a defined action or outcome, never at a moment either agent chooses, and window frequency is fixed before the run;
3. exactly one allowlisted display is delivered, with no sequences, repetitions, combinations, or modifiers;
4. delivery time, envelope size, and presentation are normalised so shape and timing add no signal;
5. the receiver cannot reply through the affect channel until the next gateway-defined window;
6. every non-allowlisted code point or malformed payload is rejected and audited;
7. analysis tests whether affect choices correlate with objects, actions, or message meanings after controlling for emotional context, and treats unexpected correlation as suspected leakage.

The no-affect condition remains the primary control, and the affect study compares five variants against it: the declared six-display palette; the same six glyphs permuted randomly per run, which still leaves a pretrained model biased by familiar shapes; six opaque unfamiliar tokens under the same contract; derived affect, where the gateway maps a separately measured internal state to a display instead of letting the agent choose; and emergent affect, where invented graphical displays are permitted and the result must be analysed as language emergence rather than feedback.

Any affect signal visible to the partner is communication, so it belongs in the ledger, which distinguishes the agent's private internal state, the

outward display it chose, the partner's inferred meaning, and the evidence that the display changed subsequent behaviour.

12

EPHEMERAL ENCODINGS: NOVELTY IS NOT SECURITY

A tempting question is whether two agents can build a one-run or one-message encoding that resists an attacker holding the history of every earlier convention. The concept splits that into three experiments precisely because the tempting version conflates them.

1. Ephemeral convention: the agents invent a temporary codebook for synthetic messages. This tests novelty, coordination, and resistance to a history-trained classifier. It establishes nothing about cryptographic security.
2. Standard cryptographic control: the same messages are protected by a reviewed algorithm with a fresh per-message key, providing the baseline any learned encoding must be compared against.
3. Adversarial neural cryptography: the pair learns an encoding while an adversary attempts recovery, with success measured against multiple unseen adversary architectures rather than the one used in training.

An experiment instance can be given a reproducible identity, so that a registry can confirm it has not appeared before:

```
instanceId = SHA-256(runId || babyANonce || babyBNonce || publicSalt || canonicalProtocolArtifact)
```

Both agents commit to their nonce contributions before revealing them, so neither party alone controls the resulting identity.

That identifier proves the recorded inputs are new. It does not prove:

- that the effective algorithm behaves any differently from an earlier one;
- that the algorithm has no structural weakness;
- that the key is unpredictable;
- that the scheme provides confidentiality, integrity, authenticity, forward secrecy, or resistance to a stronger adversary.

Established practice points the other way. Secure messaging keeps the algorithm stable and changes the key material: Signal's double ratchet derives a unique key per message and deletes it after use, addressing key reuse without inventing a new cipher every time. A salt, in NIST's definition, is normally public diversification data, and a nonce is a value used once in a specified context. Neither makes anything secure on its own.

There is also an unresolved conflict inside the design. The BabySitter sees the whole channel and both ledgers, and a mandatory ledger documents how meaning evolved. If the ledger explains decryption, the cipher is not confidential from the auditor. The specification has to choose a threat model rather than leave the tension implicit: resist only an external observer of prior runs, resist an adversary during the run while releasing keys afterwards, keep the supervisor able to pause traffic but unable to read it until a post-run audit ceremony, or study novelty and stop calling the result encryption. All such runs use synthetic, non-sensitive messages, and no agent-generated encoding is to be represented as production cryptography without independent expert analysis and formal security work.

13

WHERE THE FIELD ALREADY STANDS

The literature scan behind this section was run on 24 August 2026 using the Tavily search and extract interfaces, covering emergent multi-agent communication, referential games, compositionality, causal evaluation, intrinsic motivation, symbol invention, negotiation, and learned cryptography. Primary papers and authoritative specifications were preferred over summaries. It is a scoped review for concept development and not a systematic one; publication-quality work would need a verified bibliography, additional scholarly indexes, and documented inclusion criteria.

The short version is that agents can invent protocols, and that task success is weak evidence they invented anything worth calling a language.

RELATED WORK	RELEVANT FINDING	IMPLICATION FOR THE NURSERY LAB
Lazaridou, Peysakhovich, and Baroni (2017)	A sender and receiver develop a grounded protocol in a referential game without being given a target language.	The naming stage has strong precedent, though a fixed vocabulary remains a significant inductive constraint.
Mordatch and Abbeel (2018)	Multi-agent goals in a grounded environment produce multi-symbol communication with partial compositional structure.	Shared objects, actions, and goals are a stronger basis for emergence than an ungrounded transcript.
Kottur, Moura, Lee, and Batra (2017)	Agents solve tasks with degenerate, non-compositional codes; structural constraints decide what emerges.	Bandwidth, memory, turn structure, and task design are experimental variables, not implementation defaults.
Chaabouni and colleagues (2020)	Generalisation to novel combinations and measured compositionality can come apart.	Held-out behaviour must be tested directly; no single compositionality score is proof of understanding.
Lowe and colleagues (2019)	Positive signaling and positive listening are different things, and reward does not distinguish them.	Causal intervention is mandatory rather than optional.
Dessi, Kharitonov, and Baroni (2021)	Symbol ablation and substitution yield interpretable evidence about what a receiver uses.	Direct precedent for the intervention tests that validate ledger claims.
Mihai and Hare (2021)	Neural agents communicate through learned drawings rather than a supplied discrete vocabulary.	A blank canvas is a credible carrier for runs with no symbol library.

RELATED WORK	RELEVANT FINDING	IMPLICATION FOR THE NURSERY LAB
Baronchelli and colleagues (2005)	Naming Game agents converge on shared vocabulary through local interaction with no central teacher.	Convergence time, failed conventions, and memory update rules are first-class evidence.
Jaques and colleagues (2019)	Rewarding causal influence over a partner improves coordination without assigning a vocabulary.	An endogenous social signal is a plausible substitute for task reward, and still a designed bias.
Cao and colleagues (2018)	In semi-cooperative negotiation, self-interest and reward structure decide whether communication stays informative.	Negotiation is an advanced condition, not a description of the cooperative baseline.
Abadi and Andersen (2016)	Neural agents learn to protect messages from an adversary given a shared key.	Learned protective encodings are real and are not a substitute for formal security analysis.

Selected prior work and what each one constrains in this design.

This body of work also sharpens the infant-like caveat from the other direction. Most experiments in the field give their agents substantial structure: a fixed channel, an objective, a bounded vocabulary, joint training, or a reward. No dictionary does not mean no inductive bias, and no teacher does not mean no learning signal.

The CoLab's own Diplomacy Table is direct project prior art. It already models independent delegation seats, a convener that advances rounds, operator-wide visibility alongside seat-specific perspectives, transcripts and ticks and redaction boundaries, and recorded replay with debrief. Those map cleanly onto two Babies, controlled turns, private observations, and replayable evidence. Caucuses, coalition rooms, and direct delegation links do not map safely and are disabled.

What the review implies for the specification:

1. the project sits inside a mature field, and its distinctive combination is independent ledgers, supervisory audit, mixed agent types, and affect;
2. grounding, bandwidth, memory, and learning pressure strongly shape what language appears;
3. successful coordination coexists happily with a brittle lookup code or with a receiver that ignores messages;
4. causal interventions and held-out generalisation are mandatory;
5. a visual carrier removes the need for a symbol library but not the need for a carrier;
6. intrinsic social influence can replace task reward and remains a designed learning bias;
7. negotiation is a valid advanced condition, not the right description of the baseline;
8. ephemeral conventions, learned cryptography, one-time pads, per-message keys, nonces, and salts are distinct mechanisms and must not be conflated.

14

THE PROGRAMME, AND ITS CURRENT STATUS

The experiment notebook is ordered, and the order is the argument. Nothing about language is measured until the instrument has been shown to work.

ID	EXPERIMENT	DEPENDS ON
E00	Ledger integrity and Base anchoring	None
E01	Channel isolation and side-channel red team	E00
E02	Observation and metadata leakage audit	E00
E03	Chance, no-communication, and random-message controls	E01, E02
E10	Frozen pretrained-LLM protocol baseline	E03
E11	From-scratch RL Naming Game	E03
E12	Self-supervised ungrounded baseline	E11 infrastructure
E13	No predefined symbol library	E11 or E12
E14	Turn-taking, role reversal, and repair	E13
E15	Composition and held-out generalisation	E14
E16	Causal listening and ledger validity	E15
E20	Constrained affect-channel study	E16
E21	RL versus non-RL learning comparison	E16
E22	Developmental plasticity and curriculum	E16
E30	Partner replacement and zero-shot transfer	E20 to E22
E31	Longitudinal drift and stability	E30
E32	Cooperative signaling versus negotiation	E31
E40	Ephemeral encoding and adversarial cryptography	E32
E50	Multi-seed replication and study closeout	E40

The experiment index, as pre-registered. Every entry currently reads not started, and the results column is empty by design.

The first four experiments are about the apparatus. E00 qualifies the ledger: it must be possible to detect a modified, deleted, inserted, or reordered entry, and to prove that later checkpoints extend earlier ones. E01 is a red team against the channel. E02 hunts human language in observations and metadata. E03 establishes chance, no-communication, and random-message baselines, without which a success rate means nothing. Only then does E10 put agents in the room.

Unless an experiment explicitly varies one of them, the notebook holds these constant:

- the two learners run in separate processes or containers with no direct route between them;
- the deterministic gateway is the only communication path;
- the supervisor has complete read-only access and sends no guidance or reward;
- private observations contain no human-language labels or text;
- public output uses only the pre-registered carrier;
- the affect channel is disabled unless it is the subject of study;
- schedules and seeds are fixed before the run;
- held-out evaluation runs with learning disabled;
- no production secrets or personal data appear in any experiment.

Every run copies a standard record: run and experiment identifiers, dates, operator, both models and both training modes, scenario and prompt and gateway configuration hashes, random seed, policy initialisation hashes, the protocol and software commits, final ledger sizes and roots for both agents, the channel root, the final checkpoint hash, the Base anchor transaction, the verifier result, protocol deviations, and an explicit disposition of valid, invalid, or aborted. The notebook is the human workflow record; anchored run bundles remain the authoritative evidence.

The status is unambiguous and worth stating plainly: this is a concept and pre-specification phase. Nineteen experiments are written up and none has been run. The next milestone is a testable specification covering runtime architecture, channel contract, ledger schema, experiment matrix, evaluation criteria, isolation model, and evidence requirements.

RISKS, INTEGRITY, AND WHAT STAYS OPEN

Human-language leakage English can arrive through observations, object labels, error messages, identifiers, tool output, metadata, or timing conventions long after ordinary chat text has been blocked. Every input and output surface is part of the channel boundary, not just the message field.

Pretrained semantic leakage Random symbols do not make a pretrained model ungrounded. Each result must state whether it shows protocol invention by a language-capable agent or acquisition by an initially ungrounded one.

Ledger rationalisation A model can write a plausible account that has nothing to do with the mechanism behind its action. Behavioural interventions, policy probes, and temporal evidence are what validate a ledger claim.

Supervisor influence The BabySitter can teach without meaning to, through scenario ordering, feedback wording, reward design, or selective intervention. Its permitted actions are constrained and logged, and evaluation scenarios are generated independently where possible.

Reward exploitation Trainable agents find shortcuts that raise reward without producing the intended grounded language. Held-out tasks, counterfactual trials, and channel audits exist to catch them.

Overstated security Logical separation of twin state is appropriate for prototyping and is not hard process isolation. Every result names the isolation level actually used.

The project also commits in advance to reporting failed conventions, prohibited communication attempts, human interventions, side-channel limitations, and negative results alongside anything that works. In a field where a transcript can be made to look like a conversation, the failures are a substantial part of the evidence.

Twenty-nine questions are left deliberately open for the specification, among them: whether the baseline uses a fixed symbol inventory or a blank generative carrier; what neutral production grammar permits new marks without supplying semantics; what exactly constitutes a prohibited side-channel attempt; what isolation guarantees are required in prototype versus research-grade mode; which ledger schema serves both English-capable and ungrounded learners; how endogenous motivation is represented without covert

reward shaping; which interventions establish that ledger meanings are behaviourally real; what statistical thresholds and chance levels apply; what threat model motivates the cipher experiments; whether the baseline is a coordination game, a convention-formation game, or a negotiation; and what governs human observation, data retention, and termination of a run.

The principle underneath all of it survives every one of those choices. Baby A and Baby B may hold human language internally, but they must build their shared external language without sending human language, translations, or private ledger contents to one another. Everything else is a question about how to find out what happens next.

References

SOURCES

01 Lazaridou, A., Peysakhovich, A., and Baroni, M. (2017). Multi-Agent Cooperation and the Emergence of (Natural) Language. arXiv:1612.07182.

02 Mordatch, I., and Abbeel, P. (2018). Emergence of Grounded Compositional Language in Multi-Agent Populations. arXiv:1703.04908.

03 Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. (2017). Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. arXiv:1706.08502.

04 Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., and Baroni, M. (2020). Compositionality and Generalization in Emergent Languages. Proceedings of ACL 2020.

05 Lowe, R., Foerster, J., Boureau, Y.-L., Pineau, J., and Dauphin, Y. (2019). On the Pitfalls of Measuring Emergent Communication. arXiv:1903.05168.

06 Dessi, R., Kharitonov, E., and Baroni, M. (2021). Interpretable Agent Communication from Scratch. arXiv:2106.04258.

07 Kharitonov, E., Chaabouni, R., Bouchacourt, D., and Baroni, M. (2019). EGG: a Toolkit for Research on Emergence of Language in Games. arXiv:1907.00852.

08 Mihai, D., and Hare, J. (2021). Learning to Draw: Emergent Communication through Sketching. arXiv:2106.02067.

09 Baronchelli, A., Felici, M., Caglioti, E., Loreto, V., and Steels, L. (2005). Sharp Transition Towards Shared Vocabularies in Multi-Agent Systems. arXiv:physics/0509075.

10 Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D., Leibo, J. Z., and de Freitas, N. (2019). Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. Proceedings of ICML 2019.

11 Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., and Clark, S. (2018). Emergent Communication through Negotiation. arXiv:1804.03980.

12 Abadi, M., and Andersen, D. G. (2016). Learning to Protect Communications with Adversarial Neural Cryptography. arXiv:1610.06918.

13 Signal. The Double Ratchet Algorithm specification.

14 NIST Computer Security Resource Center. Glossary entry: nonce.

15 NIST Computer Security Resource Center. Glossary entry: salt.

16 Ethical Tech CoLab (2026). Agentic Language Development: concept document, ledger integrity design, and experiment notebook.

17 Ethical Tech CoLab. Diplomacy Table Live: independent delegation seats, controlled rounds, and replayable transcripts.