

Compositionality Needs More Than Iterated Learning: Counterfactual Pressure Makes Grammar Emerge as Operators Stable Under Intervention

Cercle Digital (independent research)
Correspondence: research@cercle.digital

Abstract

Why do learned communication codes usually fail to generalize compositionally? We test the two dominant hypotheses head to head, under pre-registered protocols with seed-resampled confidence intervals. **Iterated learning is a real but weak, slow and fragile lever.** Iterated learning (culturally transmitting the code through a learning bottleneck across generations of fresh learners, after Kirby 2001) fails to produce zero-shot composition in four pre-registered variants (a capacity bottleneck, a continuous transmitted code, a discrete transmitted code, transmitted grounded plans), each null carrying a diagnosed mechanism. Canonical neural iterated learning, run at its most favourable settings on the very world of our positive result, does lift zero-shot composition, and we report it: $0.048 \rightarrow 0.653 \pm 0.134$ over ten generations at the widest bottleneck, ending above its no-bottleneck control (0.444); but two thirds of the lift is relearning by fresh agents rather than the bottleneck, and ten generations of retraining top out below what one generation of the following pressure delivers. **Counterfactual pressure is the dominant lever.** Requiring that when one latent factor changes under an intervention ($\text{do}()$, acting on the world and reading the reproducible response), only the corresponding message position changes, with the position-to-factor binding *chosen* by the code rather than given, a two-symbol channel reaches zero-shot composition 0.865 ± 0.155 on held-out factor combinations (chance 0.062; the same channel under listener success alone scores seen pairs at 0.999 yet transfers 0.083), with topographic similarity 0.960 (meaning distance predicts message distance). The same lever operates upstream of language: on a basis recovered by a causal admission gate (axes aligned by intervention responses and purified by cross-view invariance) a compositional rule generalises at 0.935 versus 0.780 entangled, and a relation recovered under intervention chains to unseen depths (solve rate flat at 1.00 for depths 1–5) while passive clustering compounds its type error ($\approx 0.79^d$). It also operates downstream, on grounded naming: an emergent word transfers to an unseen observation map only when perception shares an invariant basis (0.728 out-of-distribution versus 0.218 adversarial), while grammar never emerges for free from naming games (zero-shot 0.000). A pilot with three factors including a relation and hard distractors scales the lever (1.000 versus free 0.325). We argue grammar is best understood as a code of *operators stable under intervention*, and note that this is precisely the axis (compositional out-of-distribution transfer) where next-token models are weakest. A self-contained code supplement reproduces every number.

1 Introduction

When two neural agents invent a discrete code to solve a reference task, the code that emerges is almost always the wrong kind of code: it names every situation it has seen and none it has not. Communication succeeds, memorisation wins, and the signature capacity of natural language, describing a *new* combination of familiar parts with familiar words, does not appear (Kottur et al., 2017; Lazaridou et al., 2018; Chaabouni et al., 2020). The question this paper measures is: *what pressure turns a discrete channel into a compositional one?*

Two families of answers dominate. The first, from language evolution, is **cultural transmission**: compositionality is the attractor of *iterated learning*, the repeated relearning of the code by fresh learners through a transmission bottleneck; only compressible (systematic) mappings survive the bottleneck, so structure accumulates across generations

(Kirby, 2001; Kirby et al., 2008; Ren et al., 2020). The second is **structural pressure**: something about the task, the input, or the environment must force the factorisation (Lazaridou et al., 2018; Hill et al., 2020; Rita et al., 2022). The two families are rarely put on the same world with the same seeds and the same pre-registered gates. We do that here, and we propose and measure a specific structural pressure that the emergent-communication literature has not isolated: *counterfactual modularity*, the requirement that an intervention on one latent factor of the meaning move exactly one part of the message.

Terminology. Five terms recur from the title onward. *Iterated learning* (Kirby, 2001) is the cultural-transmission loop above; its *transmission bottleneck* is the restriction that each new generation learns the code from a subset of the previous generation’s productions. $\text{do}()$ denotes a Pearl-style *intervention* (Pearl, 2009): setting one latent factor of the

world (or of a copy of it) and reading the reproducible response, as opposed to passively observing correlations. A *referential game* (Lazaridou et al., 2017) is the standard emergent-communication task: a speaker sees a meaning, emits discrete symbols, and a listener must recover the meaning. *Zero-shot (combinatorial) composition* is accuracy on held-out combinations of factor values, each value having been seen in training with other partners; it is the falsifiable signature of grammar, as opposed to accuracy on seen pairs, which memorisation also delivers. *Topographic similarity* (Brighton & Kirby, 2006) is the rank correlation between pairwise distances in meaning space and in message space: a structured code maps similar meanings to similar messages.

The claim. Our results support one sentence:

Grammar is a code of operators stable under intervention: a channel becomes compositional when changing one factor of the world must change one operator of the message, and the same controllable-and-invariant geometry that factors a representation upstream is what makes a grounded word transfer downstream.

Concretely, the paper establishes five measured findings.

1. **Four transmission-based attempts fail, each with a diagnosed mechanism (§4).** A capacity bottleneck fails because a pre-factored input pre-solves the task; an iterated chain with a continuous code fails on a learnability dichotomy (the entangled input is either still linearly factored or unlearnable); a discrete transmitted code drifts without composing; and transmitting grounded plan demonstrations maintains, but never grows, a typed theory. These are honest nulls with confidence intervals, not strawmen: each variant was pre-registered with the control that could have made it succeed.
2. **Head to head on one world, canonical iterated learning at its most favourable settings lifts composition partially, and does not close the gap (§5).** With imitation-then-interaction chains (Ren et al., 2020) and swept bottleneck sizes, zero-shot rises across generations at every bottleneck ($0.048 \rightarrow 0.653 \pm 0.134$ at the widest), and at that setting the final generation is CI-above the no-bottleneck control ($+0.209 \pm 0.104$): on this world the bottleneck is a real pressure, which we report. But the control itself reaches 0.444 (relearning by fresh agents does most of the work, consistent with Li & Bowling 2019), the effect needs ten generations of full retraining, and it remains below what the counterfactual pressure reaches in one.
3. **Counterfactual pressure makes the same channel compositional (§6).** With the modularity loss of Eq. 1, and the position-to-factor binding chosen by the code itself, zero-shot composition reaches 0.865 ± 0.155 (chance

0.062 ; success-only channel 0.083), topographic similarity 0.960 ± 0.054 , and the decoded command selects the correct unseen referent without exhaustive search (0.865).

4. **The lever is not specific to language (§7).** Upstream, on representations: a basis whose axes are aligned by intervention responses and purified by cross-view invariance makes a compositional rule generalise at 0.935 ± 0.024 versus 0.780 ± 0.034 on the raw entangled view, and a type-level relation recovered under intervention chains to depths never demonstrated (flat 1.00) while passive clustering decays as $\approx 0.79^d$.
5. **Downstream, on grounded naming, the same geometry governs transfer (§8).** An emergent word transfers to an unseen observation map exactly when the perception it is grounded on shares an invariant basis across maps (0.728 ± 0.072 versus 0.218 ± 0.081 adversarial), and grammar does not come free from the naming game even on that invariant grounding (zero-shot 0.000 ± 0.001). A three-factor pilot with a relation and hard distractors extends the counterfactual lever (1.000 versus 0.325; §9).

Every experiment is pre-registered (hypotheses and success criteria fixed before the run, printed by the bench script before its verdict), run over ≥ 10 seeds (12 for the channel benches; the pilot uses 8 and is labelled a pilot) with 95% confidence intervals, and reproducible from a self-contained code supplement (Appendix A).

2 Related work

Iterated learning. The iterated-learning programme showed, first in simulation (Kirby, 2001) and then with human participants (Kirby et al., 2008), that compositional structure can accumulate when a language is repeatedly transmitted through a learning bottleneck. Neural instantiations followed: Ren et al. (2020) alternate an imitation (learning) phase and an interaction phase and report higher topographic similarity and generalisation; Guo et al. (2019) obtain number codes whose structure depends heavily on the input representation; Li & Bowling (2019) show that merely *resetting* listeners (ease-of-teaching) structures the code, without a transmission chain in the cultural sense. Our contribution to this line is not a refutation of iterated learning as a model of cultural evolution; it is a measurement: on worlds where memorisation is the cheap default, four transmission-based pressures fail to produce zero-shot composition, and the canonical neural variant, at its most favourable settings on our main world, lifts composition slowly and partially, with most of the lift owed to relearning (consistent with Li & Bowling 2019) and a minority share to the bottleneck at its widest setting.

Emergent communication. Discrete codes emerge readily in referential games (Lazaridou et al., 2017; Havrylov & Titov, 2017), but are neither natural-language-like (Kotzur et al., 2017) nor compositional by default; Chaabouni et al. (2020) show at scale that emergent languages generalise without being compositional, and that compositionality is not *necessary* for generalisation. Our claim is calibrated against theirs: we measure zero-shot *combinatorial* transfer on held-out factor combinations under a fixed small channel, the regime where Auersperger & Pecina (2022) argue compositionality genuinely matters, not generalisation in the loose sense. Environmental and objective-side pressures have been studied (Lazaridou et al., 2018; Hill et al., 2020; Rita et al., 2022); the specific pressure we isolate, one do() flip moving one message position with the binding left free, has to our knowledge not been tested as the sole structural addition to an otherwise standard game.

Compositional generalisation and disentanglement. Zero-shot recombination is exactly where sequence models are weakest (Lake & Baroni, 2018; Keysers et al., 2020), a weakness that persists in large transformer models on compositional tasks (Dziri et al., 2023); taxonomies are given by Hupkes et al. (2020) and measurement frameworks by Andreas (2019). On the representation side, Higgins et al. (2018) ground compositional concepts in disentangled visual primitives, while Montero et al. (2021) report that disentanglement per se does not guarantee combinatorial generalisation in generative models. Our upstream result agrees with the nuance: what buys composition in our benches is not disentanglement as a statistical property of a passively trained encoder, but a basis whose axes are *defined* by intervention responses and cross-view invariance, the two signals causal representation learning identifies as foundational (Schölkopf et al., 2021). Grounding words in such representations connects to the symbol-grounding problem (Harnad, 1990).

3 Common machinery and metrology

3.1 Worlds and channels

All benches share three design decisions, chosen so that memorisation is the cheap default and composition has to be *caused* by the pressure under test.

Holistic rendering. Meanings are never given to the learner pre-factored. A meaning with factor vector f (for example, colour and shape one-hots) is rendered through a fixed random nonlinear mix, $x = \tanh(Wf) + \varepsilon$, so the factorisation must be discovered. Where a bench needs the pre-factored input as a control (the capacity null), that is stated explicitly.

Held-out combinations. Train and test splits are by *combination*: every factor value appears in training with other

partners, and the test set contains only unseen combinations (for the 4×4 channel worlds, train off-diagonal, test on the four diagonal pairs; both-correct chance $1/16 = 0.062$).

Discrete channels. Channel benches use Gumbel-softmax straight-through speakers (Havrylov & Titov, 2017) with message length L and vocabulary 6, and listeners with one head per factor. The reasoning benches (§4, §7) use the same compositional target throughout: a meaning is a pair (obstacle type o , present set S of tool types), and the label `craft?` = $[\text{needed}(o) \notin S]$ composes a learned unary relation ($o \mapsto \text{needed}$) with a membership test; a factored representation computes it on unseen (o, S) pairs, a holistic one memorises.

3.2 Metrology

Each bench fixes its seeds explicitly (a run is a pure function of (configuration, seed)), uses ≥ 10 seeds (12 for the channel benches; 8 for the pilot, labelled as such), and reports mean $\pm$ half-width of the 95% interval (1.96 SE), written 0.865 ± 0.155 . Hypotheses and success criteria are pre-registered in the bench scripts, which print them together with their controls before the verdict; a hypothesis fails when its confidence interval crosses the criterion, and failed hypotheses are reported as results with their diagnosed mechanism, not reframed. Every claimed separation is CI-separated (the interval of the paired per-seed difference excludes zero). Controls are strong by construction: every positive arm has a matched arm that differs by exactly the pressure under test, and every null has the control that could have made it succeed.

4 Transmission pressures do not induce composition: four nulls

This section reports four pre-registered attempts to make composition *emerge* from compressibility-based pressures, in increasing order of fidelity to the iterated-learning recipe. All four fail. Their value is in the mechanisms the controls expose, which is why we report them in full (Table 1).

4.1 A capacity bottleneck is not a transmission bottleneck

Hypothesis (pre-registered). If compressibility is the pressure, a tight capacity bottleneck k on the learned representation (near the number of types $M=7$) should force factoring into reusable atoms and yield *higher* zero-shot accuracy on held-out (o, S) combinations than a loose one, which is free to memorise (CAPACITY-BOTTLENECK).

Experiment (operationalisation). The `craft?` task of §3.1 with the input given *pre-factored* (o one-hot concate-

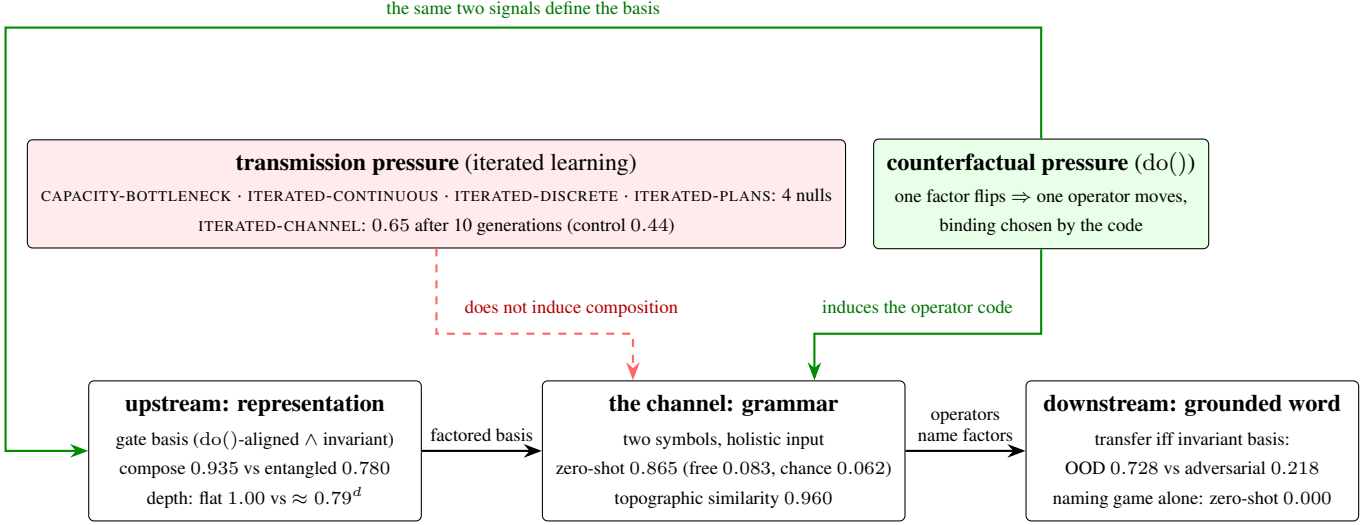


Figure 1: One lever at three levels. Two candidate pressures compete for making a discrete channel compositional. The transmission pressure (left, red) fails in four pre-registered variants and, in its canonical neural form at its most favourable settings, needs ten generations to reach two thirds of what one generation of the other pressure delivers (§4, §5). The counterfactual pressure (right, green) makes the channel an operator code (§6); the same two causal signals (intervention response, cross-view invariance) factor the representation upstream (§7) and govern whether a grounded word transfers downstream (§8). All numbers are means over ≥ 10 seeds; benches named in small capitals are scripts of the code supplement.

nated with S multi-hot), an encoder squeezed to width $k \in \{2, 3, 4, 6, 10, 24\}$, 55/45 train/held-out split by combination, 10 seeds.

Result. The hypothesis fails in both directions: held-out accuracy *rises* with capacity (0.517 ± 0.020 at $k=2$, 0.958 ± 0.023 at $k=6$, 0.970 ± 0.018 at $k=24$) and the high-capacity learner does not memorise (train 1.000, held-out 0.970).

Conclusion. The null has a design diagnosis that shapes everything after it: the *pre-factored input pre-solves the task*. A plain network composes at any width because the atoms are handed to it; there is no memorisation to break, hence no work for a pressure to do. Any honest emergence test must render the meaning holistically, which all subsequent benches do.

4.2 Iterated learning of a continuous code: a learnability dichotomy

Hypothesis (pre-registered). With the input entangled and the canonical chain in place (each generation learns from a bottlenecked 30% sample of the previous generation’s productions; a grounding anchor re-checks half of each sample against the true rule, supplying the expressivity pressure that keeps the chain from playing telephone), held-out accuracy should *rise* across 12 generations and end

CI-above both the transmission-only arm (drift control) and the no-transmission arm (ITERATED-CONTINUOUS).

Experiment (operationalisation). Four arms (full recipe; transmission-only; no-transmission control; pre-factored control) crossed with two entangling regimes: a weak mix (tanh of a random linear map) and a strong one (random Fourier features), 10 seeds, transmitted code = the learner’s 6-dimensional latent message.

Result. Null in both regimes, and the two failures are opposite. Under the weak mix the task is solvable holistically within one generation (full recipe 0.81 ± 0.03 at generation 0) but nothing rises: the final generation is *lower* (0.77 ± 0.02 ; rise -0.05 ± 0.04), transmission alone decays ($0.79 \rightarrow 0.57$: the chain plays telephone and the grounding anchor merely rescues it, gap $+0.20 \pm 0.04$ over drift), the no-transmission control is *above* the full recipe (0.85 ± 0.02), and neither topographic similarity nor the type decode rises (the type decode is already 0.99 at generation 0: the factoring is present before any transmission happens). Under the strong mix everything sits at chance (0.52 ± 0.03 flat, topographic similarity ≈ 0) while the pre-factored control solves the same task (0.80 ± 0.03): the task is fine, the holistic version is unlearnable at this scale.

Conclusion. The result is a *learnability dichotomy*, not an emergence: under the weak mix the type axes remain effectively linearly decodable, so the factoring is present at gener-

ation 0 and there is no compressibility gradient for the chain to climb (the anchor merely rescues the chain from drift); under the strong mix the task is unlearnable within one generation even pre-factored into the network, so there is nothing for the chain to refine. The binding variable is where the input sits on learnability, not the transmission pressure. This diagnosis motivates the discrete channel next, since a continuous message can leak holistic information through the bottleneck.

4.3 A discrete transmitted code drifts without composing

Hypothesis (pre-registered). With the two ingredients the previous null lacked, a *discrete* two-symbol channel (holistic codes cannot pass a narrow symbolic bottleneck unchanged) and iterated transmission of the discrete messages themselves, held-out accuracy under the full recipe should rise across 10 generations and end CI-above the no-transmission and single-symbol controls, and topographic similarity should rise (ITERATED-DISCRETE).

Experiment (operationalisation). Speaker imitates the previous generation’s discrete messages on the bottleneck sample (cross-entropy on tokens); listener grounded on half the sample by the true rule; controls: no-transmission, and a single-symbol channel with vocabulary smaller than the meaning space (structure floor); 10 seeds.

Result. Null on both hypotheses. Held-out accuracy under the full recipe is flat across generations (0.64 ± 0.04 at generation 0, 0.62 ± 0.03 at generation 9; rise -0.02 ± 0.04), does not separate from the no-transmission control (gap -0.03 ± 0.05), and exceeds only the single-symbol floor ($+0.07 \pm 0.04$), an effect of channel capacity, not of transmission. Topographic similarity does not rise ($+0.03 \pm 0.05$, interval crossing zero).

Conclusion. The discrete bottleneck transmits, but what survives it is not structure: the chain neither gains nor loses what the grounding anchor puts in each generation.

4.4 Transmitting grounded plans maintains a theory, never grows one

Hypothesis (pre-registered). The strongest transmission object: (situation $\rightarrow$ verified plan) demonstrations over an affordance world, with each generation inducing a typed theory (cluster entangled percepts into types, read type-level rules off the received demonstrations, re-ground half against the live world) and acting with a shared backward-chaining planner. Under the bottleneck, novel-instance transfer and extrapolation to depth 2 (never demonstrated) should rise across 12 generations; the no-bottleneck control must be

able to fail, and must fail, for a rise to count (ITERATED-PLANS).

Experiment (operationalisation). Four arms (bottleneck; no-bottleneck control; typed ceiling with full live grounding; flat object-identity floor); 10 seeds.

Result. Null, with an unusually clean mechanism. The no-bottleneck control coincides with the typed ceiling (0.90 ± 0.20 novel transfer, 0.80 ± 0.26 depth-2, flat across generations): given full demonstrations, the induced typed theory is already the best theory at generation 0. The bottleneck arm does not rise toward it; it sits low and flat (0.20 ± 0.09 at generation 0, 0.17 ± 0.08 at generation 11; rise -0.03 ± 0.14 ; depth rise $+0.06 \pm 0.08$, interval crossing zero), because the instance bottleneck starves the type clustering itself. The floor (object-identity memorisation) is 0.00 everywhere.

Conclusion. Together the four nulls triangulate one lesson: in every variant, whatever composition exists is *put in* by grounding or by the input representation, and the transmission dynamics at best preserve it. Nothing in the chain manufactures structure. The pressure has to come from somewhere else.

5 Head to head: the strongest iterated arm on the world of the positive result

A fair reader of §4 can object that each null might be unfavourable to iterated learning in its own way: wrong transmitted object, wrong learnability regime. ITERATED-CHANNEL removes the objection by running *canonical* neural iterated learning (Ren et al., 2020) on exactly the world, channel, split and seeds of the positive result in §6, tilted in its favour wherever a choice existed.

Hypothesis (pre-registered). If the transmission bottleneck is the lever, zero-shot composition should rise across generations under some bottleneck size, end CI-above generation 0 *and* CI-above the no-bottleneck control, and approach what the counterfactual pressure achieves.

Experiment (operationalisation). Chains of 10 generations of fresh speaker/listener pairs. Each generation first *imitates* the previous generation’s greedy messages on a transmission-bottlenecked subset of the 12 seen meanings (sizes 6, 8, 10 swept; 12 = the no-bottleneck control), then trains on communication success over *all* seen meanings (the success pressure is never handicapped). Every generation is scored and the *best* generation is reported alongside the final

Table 1: Four pre-registered transmission-based attempts and their diagnosed mechanisms. Each arm had the control that could have made it succeed; “rise” means a CI-separated increase from generation 0 to the final generation. Full curves in the bench logs of the code supplement.

Bench	Object transmitted	Pressure realised	Key control	Outcome
CAPACITY-BOTTLENECK	(none: capacity squeeze)	compressibility of the representation	pre-factored input	held-out <i>rises</i> with capacity; 0.958 at $k=6$ vs 0.970 at $k=24$
ITERATED-CONTINUOUS	continuous latent code	bottleneck + grounding anchor	no-transmission; two entangling regimes	learnability dichotomy: weak mix pre-factored (decode 0.99 at gen 0, no rise), strong mix at chance flat (0.64 $\rightarrow$ 0.62); no rise vs control (-0.03 ± 0.05)
ITERATED-DISCRETE	discrete 2-symbol messages	discrete bottleneck + grounding	no-transmission; 1-symbol floor	bottleneck low and flat (0.20 $\rightarrow$ 0.17); no-bottleneck = ceiling (0.90) at gen 0
ITERATED-PLANS	verified plan demonstrations	instance bottleneck + live re-grounding	no-bottleneck (must fail)	

one (protecting the hypothesis against late-chain drift); 12 seeds, identical world and split as §6.

Result. Fig. 2. Zero-shot rises across generations at every bottleneck: from 0.048 ± 0.036 at generation 0 to 0.544 ± 0.104 (bottleneck 6), 0.576 ± 0.129 (8) and 0.653 ± 0.134 (10) at generation 9, with topographic similarity rising $0.561 \rightarrow 0.852 \pm 0.066$ under the bottlenecks. The pre-registered criterion passes at the widest bottleneck only: its final generation is CI-above the no-bottleneck control ($+0.209 \pm 0.104$); at bottlenecks 6 and 8 the gap to the control does not separate ($+0.100 \pm 0.129$, $+0.133 \pm 0.159$). The control itself rises to 0.444 ± 0.122 (topographic similarity 0.699): relearning by fresh agents, with no transmission restriction at all, accounts for about two thirds of the best iterated lift. Scoring at the best generation instead of the last changes nothing material (best 0.653, at generation 9).

Conclusion. Honestly read, this is a partial *positive* for iterated learning, and the strongest transmission result in this paper: on a world with a holistic rendering and a small channel, canonical neural iterated learning lifts zero-shot composition far above the non-iterated free channel (0.083), and the widest bottleneck contributes beyond mere relearning. Three facts calibrate it. First, decomposition: most of the lift (0.444 of 0.653) appears in the control that transmits *everything*, so the operative ingredient is mostly the relearning pressure of fresh agents (Li & Bowling, 2019), with the bottleneck adding a CI-separable but minority share ($+0.209$). Second, cost: the lift requires ten generations of full re-training of both agents. Third, ceiling: after all that, the chain’s mean remains 0.212 below what the counterfactual pressure produces in a single generation on the same seeds (0.865 ± 0.155 , topographic similarity 0.960 versus 0.852); the per-seed intervals of the two pressures overlap (both are seed-variable), so we state the comparison as it is: one generation of counterfactual pressure buys more than ten generations of the most favourable transmission chain we could build, and the four nulls of §4 show how narrow the regime is in which transmission delivers even this much.

6 Counterfactual pressure makes the channel an operator code

Hypothesis (pre-registered). If the channel is required to be *counterfactually modular*, one factor flipped under

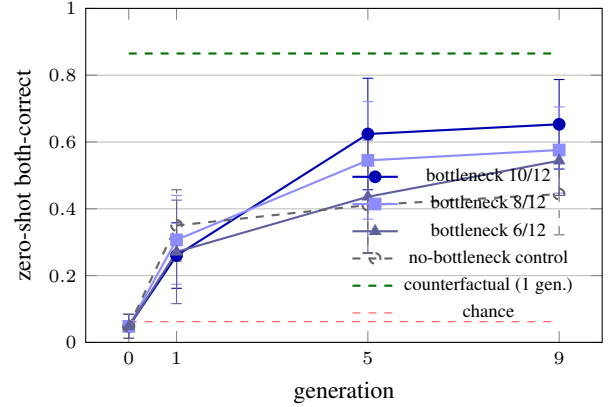


Figure 2: ITERATED-CHANNEL: canonical neural iterated learning on the world of the positive result (12 seeds, 95% CI; generations 0, 1, 5, 9 shown). Zero-shot rises at every bottleneck; only the widest bottleneck ends CI-above the no-bottleneck control ($+0.209 \pm 0.104$), and the relearning-only control reaches two thirds of the best iterated level. The green dashed line is what the counterfactual pressure reaches in a single generation on the same seeds (0.865).

do() moves exactly one message position, with the position-to-factor binding chosen by the code, then the same two-symbol channel that memorises under success pressure should compose zero-shot: held-out accuracy CI-above chance and CI-above the free channel (H1), and the decoded command should select the correct unseen referent without exhaustive search (H2) (INTERVENTIONAL-CHANNEL).

Experiment (operationalisation). Meanings are (colour, shape) pairs, 4×4 , rendered holistically (§3.1); train off-diagonal, test zero-shot on the diagonal; 900 steps, 12 seeds. Three arms: *holistic* ($L=1$, success pressure only: no syntactic degrees of freedom), *free* ($L=2$, success pressure only: the memorisation control), and *interventional* ($L=2$, success pressure plus the modularity loss). For a batch of meanings (c, s), the speaker’s per-position symbol distributions are read under three renderings: unchanged, colour flipped ($c' \neq c$, shape held), and shape flipped. For a candidate binding of positions to factors (σ), the loss demands *invariance* of a position to the factor it is not bound to and *sensitivity* (a Jensen-Shannon divergence margin $\tau = 0.45$) to the factor

Table 2: INTERVENTIONAL-CHANNEL: mean $\pm$ 95% CI over 12 seeds. Both-correct chance = 0.062.

Arm	Seen	Zero-shot	Action	Topsim
holistic ($L=1$)	0.489 $\pm$ 0.008	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000	0.475 $\pm$ 0.028
free ($L=2$)	0.999 $\pm$ 0.000	0.083 $\pm$ 0.049	0.079 $\pm$ 0.047	0.578 $\pm$ 0.040
interventional	0.976 $\pm$ 0.031	0.865 $\pm$ 0.155	0.865 $\pm$ 0.155	0.960 $\pm$ 0.054

it is bound to:

$$\mathcal{L}_{cf} = \min_{\sigma \in \mathbb{S}_2} \sum_{k \in \{c, s\}} \left[\text{KL}(p_{\sigma(k)} \| p_{\sigma(k)}^{\text{flip}(\bar{k})}) + (\tau - \text{JS}(p_{\sigma(k)}, p_{\sigma(k)}^{\text{flip}(\bar{k})}))_+ \right], \quad (1)$$

where p_j is the symbol distribution at position j , $\text{flip}(k)$ denotes the rendering with factor k intervened on, $\bar{k}$ the other factor, and $(\cdot)_+ = \max(0, \cdot)$. The minimum over the two bindings is the point of the design: the code is not told which position names colour; it *chooses* its own grammar. The action metric decodes the two symbols into attributes and succeeds only if they identify the held-out target exactly (a proxy for selecting it among distractors without touching every object).

Result. Table 2 and Fig. 3. The free channel is a textbook memoriser: seen 0.999 $\pm$ 0.000, zero-shot 0.083 $\pm$ 0.049 (chance 0.062), action 0.079. The holistic single symbol cannot even represent the diagonal (0.000). The interventional channel scores seen 0.976 $\pm$ 0.031 and **zero-shot** 0.865 $\pm$ 0.155, action 0.865 $\pm$ 0.155, topographic similarity 0.960 $\pm$ 0.054. The pre-registered gaps are CI-separated: zero-shot minus chance +0.803 $\pm$ 0.155; interventional minus free +0.782 $\pm$ 0.170; action gap +0.785 $\pm$ 0.169.

Conclusion. Both hypotheses pass. Capacity is not the difference (free and interventional share the channel); the world is not the difference (shared seeds and rendering); the only difference is the counterfactual constraint. What it buys is precisely the grammar signature: symbols become *operators* bound to factors, so unseen combinations are named by recombining operators, and the message is cheap to act on. The pressure does use privileged access: it knows *which* factor was flipped (though not which position should carry it). We return to this in Limitations.

7 Upstream: composition is the geometry of a causal admission gate

If grammar is a code of intervention-stable operators, the lever should not care whether the “code” is a message or a representation. This section measures the same lever one level below language, where the compositional object is a readout basis.

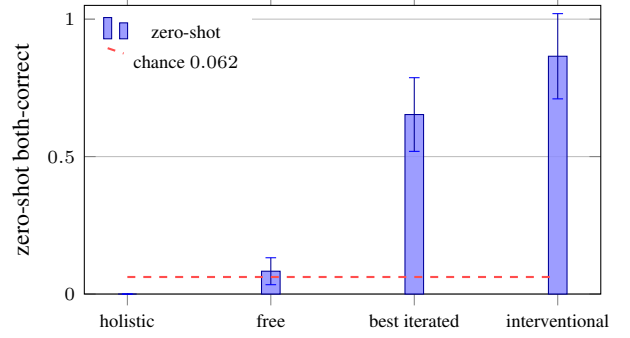


Figure 3: Zero-shot composition on held-out combinations, all pressures on the same world (12 seeds, 95% CI). Success-only channels memorise (free: seen 0.999, zero-shot 0.083); the best iterated-learning generation at the most favourable bottleneck reaches 0.653 after ten generations; the counterfactual pressure reaches 0.865 in one.

7.1 The gated basis makes composition free

Hypothesis (pre-registered). Route the entangled meaning through a basis defined by two causal signals, axes aligned by $\text{do}()$ responses (controllability) and purified by cross-view invariance, and the compositional rule should generalise zero-shot near ceiling, CI-above the entangled readout (H1), with *both* signals necessary (H2) (GATE-COMPOSE).

Experiment (operationalisation). The `craft?` task with $E=6$ holistic noisy views $x_e = \tanh(W_e f) + \varepsilon$ of the same factor vector f . The gate recovers $z_e = J_e^+(x_e - x_e^0)$, where J_e is the intervention-response Jacobian (column i = the view’s response to $\text{do}()$ setting factor i) and J_e^+ its pseudo-inverse, then averages z_e across views. Arms: raw entangled view; gate; controllability-only (single view, no cross-view purification); invariance-only (cross-view-shared PCA subspace, no intervention axes). A small MLP readout is trained on seen combinations, tested zero-shot; 12 seeds.

Result. Gate **0.935 $\pm$ 0.024** versus entangled 0.780 $\pm$ 0.034 (gap +0.16 $\pm$ 0.04, CI-separated); controllability-only 0.823 $\pm$ 0.034; invariance-only 0.812 $\pm$ 0.031 (gaps to the gate +0.11 $\pm$ 0.03 and +0.12 $\pm$ 0.03, both CI-separated). The gated basis recovers the factors at $\text{corr}(z, f) = 0.981 \pm 0.001$.

Conclusion. Composition is not learned on the gated basis; it is *inherited* from its geometry. Each signal alone is insufficient: intervention aligns axes but leaves per-view noise; invariance denoises but leaves the subspace rotated. This is the representational twin of Eq. 1: the same two requirements (respond to your own factor, be invariant to the others) that make a message position an operator make a basis vector a factor. It also sharpens the disentanglement nuance of Montero et al. (2021): what buys combinatorial

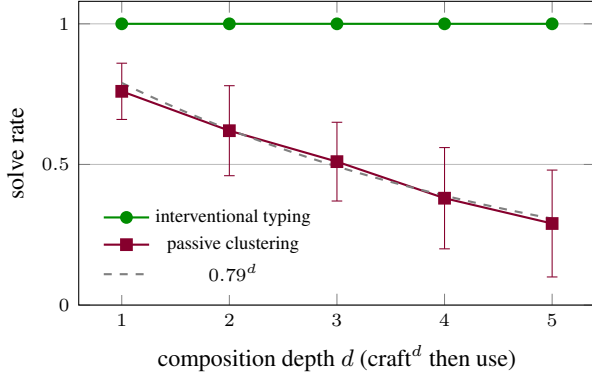


Figure 4: RECURSIVE-DEPTH: solve rate versus composition depth (10 seeds, 95% CI). The relation recovered under $\text{do}()$ chains to depths never demonstrated (flat 1.00); the passively clustered relation compounds its per-step type error, matching the geometric reference 0.79^d .

generalisation here is not a statistical disentanglement score but interventionally defined axes.

7.2 Depth extrapolation: interventional typing chains, passive typing compounds

Hypothesis (pre-registered). A type-level relation recovered *exactly* chains to any depth; one recovered with a per-item error rate p should solve depth- d problems at $\approx p^d$. Typing objects by their $\text{do}()$ response signature should therefore stay flat in depth, while typing by passive clustering of entangled percepts should decay geometrically (RECURSIVE-DEPTH).

Experiment (operationalisation). Eight types on a single-cycle craft graph (so genuine depth- d chains exist for $d < 8$); puzzles require crafting d times then using the produced tool, $d \in \{1, \dots, 5\}$; perception entangled and noisy; the same recursive backward-chaining planner consumes either typing; 10 seeds.

Result. Fig. 4. Interventional typing: solve rate 1.00 ± 0.00 at every depth. Passive typing: 0.76 ± 0.10 at depth 1 falling to 0.29 ± 0.19 at depth 5, and the decay is geometric with fitted $p = 0.79$ (the curve 0.79^d overlays the measured points).

Conclusion. Depth extrapolation is the strongest compositionality test (the code must recurse beyond anything demonstrated), and it separates the two typings maximally: an intervention-grounded operator is exact, so its compositions are exact; a passively estimated operator carries an error that *compounds* under composition. This is the mechanism behind the channel result: operators must be stable under intervention because they are meant to be *reused*.

8 Downstream: grounded words inherit the geometry, and grammar is still not free

One level above the channel, we ground naming in perception and ask two questions: what makes an emergent *word* transfer beyond its training observations, and does grammar appear once naming is grounded?

8.1 A word transfers iff its grounding is invariant

Hypothesis (pre-registered). An emergent name is a symbol bound to a region of a learned representation. It should transfer to an unseen observation map exactly when that representation does: with an observation-map family sharing an invariant basis, out-of-distribution naming should be CI-above chance and CI-above the same pipeline on an adversarial family of independent maps (WORD-TRANSFER).

Experiment (operationalisation). A latent referent (one of $M=4$, chance 0.25) is rendered through a “glass” $x = \tanh(A_e z)$; a family of nine glasses is drawn as $A_e = (1-\alpha)A_{\text{shared}} + \alpha A_e^{\text{indep}}$, with $\alpha = 0.5$ (shared structure) or $\alpha = 1.0$ (adversarial: nothing shared). An encoder is trained across eight glasses with a cross-view contrastive objective (the invariance pressure), then frozen; a speaker/listener pair plays the referential game on its output; the ninth glass, never seen by encoder or game, is the out-of-distribution test; 12 seeds.

Result. Shared family: ID 0.955 ± 0.034 , OOD 0.728 ± 0.072 . Adversarial family: ID 0.746 ± 0.054 , OOD 0.218 ± 0.081 ($\approx$ chance; OOD minus chance -0.032 ± 0.081). Shared minus adversarial OOD: $+0.510 \pm 0.102$, CI-separated; shared OOD minus chance: $+0.478 \pm 0.072$.

Conclusion. The code emerges and communicates in both regimes (H1), but its *meaning* travels only on the invariant basis. Transfer is a property of the grounding, not of the naming game: no language-side mechanism is involved in the difference between 0.728 and 0.218. This is the downstream face of §7.1: invariance decides what a symbol can be about.

8.2 Grammar does not come free from grounded naming

Hypothesis (pre-registered). Grounding might be thought to supply structure for free. On the favourable (shared, invariant) grounding of §8.1, a two-attribute scene naming game trained on seen combinations is measured for zero-shot composition; the emergent-communication

literature predicts a null (Kottur et al., 2017; Chaabouni et al., 2020), which this bench either reproduces or overturns on our grounded setting (GRAMMAR-FREE).

Experiment (operationalisation). Scenes (a_1, a_2) , two attributes $\times$ four levels, latent = sum of attribute prototypes, rendered through the shared glass family; frozen invariant encoder; $L=2$ speaker, two-head listener; train off-diagonal, test diagonal (chance 0.062); $L=1$ single-symbol control; 10 seeds.

Result. The structured code communicates seen scenes (0.546 ± 0.057 , well above chance; the $L=1$ floor scores 0.345 ± 0.028) but zero-shot composition is **null**: 0.000 ± 0.001 (zero-shot minus chance -0.062 ± 0.001), topographic similarity a modest 0.391 ± 0.048 .

Conclusion. Even with invariant grounding, success pressure rewards memorisation. Grounding fixes what symbols can refer to (§8.1); it does not force the code to factor. The pressure of Eq. 1 is doing work that neither communication success nor grounding does.

9 Pilot: three factors, a relation, hard distractors

Hypothesis (pre-registered). The announced falsification frontier for the operator-code claim is scale past two factors and past attribute-like meanings: relations, and contexts with distractors. As a first probe, with meanings (colour, shape, relation) and the binding chosen among $3! = 6$ assignments, the counterfactual pressure should still produce zero-shot composition CI-above the free channel (RELATIONS-PILOT).

Experiment (operationalisation). $4 \times 4 \times 2$ meanings (relation $\in \{\text{left-of, right-of}\}$ to a reference object), rendered holistically; $L=3$, vocabulary 6; 8 of 32 triples held out with every level seen in training; the action metric requires the decoded triple to identify the target among three hard distractors, each sharing exactly two attributes with the target; 8 seeds (pilot).

Result. Both arms saturate seen accuracy (1.000). Zero-shot: interventional 1.000 ± 0.001 versus free 0.325 ± 0.118 (gap $+0.675 \pm 0.118$, CI-separated; chance 0.031). Hard-distractor action: 1.000 ± 0.000 versus 0.320 ± 0.116 .

Conclusion. The lever survives its first scaling probe: a third factor, a relational factor among them, a six-way binding choice, and hard distractors. This is one configuration at 8 seeds and is labelled a pilot accordingly; the full falsifier

(compositional relations such as negation, real intervention budgets) remains future work.

10 Discussion

Why transmission lags where intervention succeeds. The iterated-learning argument is selectional: compressible codes survive relearning. But selection can only choose among variants the learner proposes, and a gradient-trained learner facing an entangled world proposes holistic codes first; in four of our five transmission settings the bottleneck merely preserves whatever the grounding put in (§4), and where the chain does lift composition (§5), most of the lift comes from relearning by fresh agents (consistent with Li & Bowling 2019), the bottleneck adds a minority share, slowly, and the ceiling stays low. The counterfactual pressure is not selectional but *constructive*: it names the property compositionality actually is (a factored response to interventions) and optimises it directly, while leaving the discrete choice (which position carries which factor) to the code. The chosen binding matters: it is what makes the result a small model of grammar *emergence* rather than of supervised factor labelling.

Grammar as operators stable under intervention. Across levels, the compositional objects that generalise are exactly those whose response to `do()` is sparse and reusable: a message position that moves only when its factor moves (§6), a basis vector aligned to one intervention response (§7.1), a type whose `do()` signature is exact and therefore chains (§7.2), a word grounded on an invariant basis (§8.1). This reframes grammar away from word order and toward a property that has nothing specifically linguistic about it, which is why the same lever operates on representations and plans. The natural next substrate is program code, where interventions are cheap and literal (edit one expression, rerun the tests) and counterfactual modularity is what diffs, linters and test suites verify daily.

Positioning against the compositionality debate. Chaabouni et al. (2020) show emergent languages can generalise without compositionality; our results do not contradict this, because our claim is about zero-shot *combinatorial* transfer under a small fixed channel, the regime where Auersperger & Pecina (2022) defend compositionality’s necessity, and our free channel indeed generalises nowhere (0.083) while communicating perfectly (0.999). Montero et al. (2021) show disentanglement does not guarantee combinatorial generalisation; our upstream benches agree and locate the missing ingredient: axes must be *interventionally* defined, not merely statistically independent (§7.1). Finally, next-token models are documented to be weakest precisely on compositional out-of-distribution

recombination (Lake & Baroni, 2018; Keyzers et al., 2020; Dziri et al., 2023); a pressure that buys exactly that axis, from interaction rather than from prediction, is therefore of interest beyond emergent communication.

11 Limitations

Known counterfactual axes. The pressure of Eq. 1 knows *which* factor was intervened on (it does not know, and must choose, which message position should carry it). An agent that must discover its world’s intervention axes before pressuring its channel with them is the honest next step; the upstream benches show the recovery half is feasible (the gate recovers factors at 0.981), but the two halves are not yet joined end to end. **Action is a proxy.** The action metric decodes symbols into attributes and checks identification; it is not an embodied rollout with a real intervention budget. **Synthetic worlds, small channels.** All substrates are controlled synthetic families (4×4 to $4 \times 4 \times 2$ meanings, vocabulary 6, messages of length ≤ 3); the pilot is one configuration at 8 seeds. Scale, population dynamics, and natural perceptual input are untested here. **Scope of the iterated-learning conclusion.** Our nulls and the head-to-head bound what iterated learning does *on these worlds at this scale* with gradient learners; they do not bear on its validity as a model of human cultural evolution (Kirby et al., 2008), and our own head-to-head reproduces a partial neural positive in line with Ren et al. (2020). What we claim is comparative: on a shared world with shared seeds, ten generations of the most favourable transmission chain deliver less than one generation of the counterfactual pressure, and the mean gap (+0.212) is stated with the caveat that the two pressures’ per-seed intervals overlap. **Downstream benches are simplified.** The grounded-naming pipeline uses a small contrastive encoder and linear glasses; it reproduces the qualitative regime (transfer iff shared basis) with margins large enough to survive the simplification, but richer perception is future work.

12 Conclusion

On worlds where memorisation is the cheap default, simulated cultural transmission is at best a slow and fragile source of compositionality: four pre-registered transmission pressures fail with diagnosed mechanisms, and the canonical neural chain at its most favourable settings needs ten generations to reach 0.653 ± 0.134 , two thirds of which its relearning-only control reaches without any bottleneck. A counterfactual pressure, one intervention moves one operator, with the binding chosen by the code, turns the same two-symbol channel into a compositional operator code (zero-shot 0.865 ± 0.155 versus free 0.083, topographic similarity 0.960), scales in a first pilot to three factors with a relation

(1.000 versus 0.325), and is the same geometry that makes a gated representation compose (0.935 versus 0.780), a recovered relation chain to unseen depths (1.00 flat versus 0.79^d), and a grounded word transfer (0.728 versus 0.218). Composition, on this evidence, is not a sediment of transmission; it is the shape of a code whose parts answer for their own counterfactuals. Whether the intervention axes themselves can be discovered rather than given, and whether the lever holds on natural perception and on program code, are the questions this result makes precise.

Reproducibility statement

A self-contained code supplement (the `code/` directory accompanying this paper) reproduces every number: it depends only on `numpy`, `torch` and `scikit-learn`, regenerates every world bit-for-bit from (configuration, seed), and contains one script per measured section (Appendix A). Each script prints its pre-registered hypotheses, its controls and its verdict; all reported numbers are the printed output of the full runs (10–12 seeds; the pilot at 8).

References

- Andreas, J. Measuring compositionality in representation learning. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1902.07181.
- Auersperger, M. and Pecina, P. Defending compositionality in emergent languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 285–291, 2022.
- Brighton, H. and Kirby, S. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242, 2006.
- Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., and Baroni, M. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4427–4442, 2020.
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18654.
- Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I., and Smith, K. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv preprint arXiv:1910.05291*, 2019.

- Harnad, S. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346, 1990.
- Havrylov, S. and Titov, I. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2149–2159, 2017.
- Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C. P., Bošnjak, M., Shanahan, M., Botvinick, M., Hasabis, D., and Lerchner, A. SCAN: Learning hierarchical compositional visual concepts. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1707.03389.
- Hill, F., Lampinen, A. K., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., and Santoro, A. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1910.00571.
- Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.
- Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., Tsarkov, D., Wang, X., van Zee, M., and Bousquet, O. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1912.09713.
- Kirby, S. 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, 2001.
- Kirby, S., Cornish, H., and Smith, K. 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, 2008.
- Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. Natural language does not emerge ‘naturally’ in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2962–2967, 2017.
- Lake, B. M. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2873–2882, 2018.
- Lazaridou, A., Peysakhovich, A., and Baroni, M. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1612.07182.
- Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1804.03984.
- Li, F. and Bowling, M. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15825–15835, 2019.
- Montero, M. L., Ludwig, C. J. H., Costa, R. P., Malhotra, G., and Bowers, J. The role of disentanglement in generalisation. In *International Conference on Learning Representations (ICLR)*, 2021.
- Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.
- Ren, Y., Guo, S., Labeau, M., Cohen, S. B., and Kirby, S. Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:2002.01365.
- Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., and Strub, F. Emergent communication: Generalization and overfitting in Lewis games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

A Bench index

Table 3: Bench names used in the text and the reproducing script in the code supplement.

Bench	Role in this paper	Seeds	Script
CAPACITY-BOTTLENECK	null #1: capacity squeeze; pre-factored-input diagnosis	10	bench_capacity_bottleneck.py
ITERATED-CONTINUOUS	null #2: iterated chain, continuous code; learnability dichotomy	10	bench_iterated_continuous.py
ITERATED-DISCRETE	null #3: iterated chain through a discrete channel	10	bench_iterated_discrete.py
ITERATED-PLANS	null #4: transmitted grounded plan demonstrations	10	bench_iterated_plans.py
ITERATED-CHANNEL	head-to-head: canonical neural iterated learning, favourable settings	12	bench_iterated_channel.py
INTERVENTIONAL-CHANNEL	main result: counterfactual pressure, chosen binding	12	bench_interventional_channel.py
GATE-COMPOSE	upstream: composition free on the gated basis; both signals needed	12	bench_gate_compose.py
RECURSIVE-DEPTH	upstream: depth extrapolation, flat vs geometric decay	10	bench_recursive_depth.py
WORD-TRANSFER	downstream: word transfers iff invariant grounding	12	bench_word_transfer.py
GRAMMAR-FREE	downstream: no free grammar from grounded naming	10	bench_grammar_free.py
RELATIONS-PILOT	pilot: three factors incl. a relation, hard distractors	8	bench_relations_pilot.py

B The two shared instruments

The counterfactual modularity loss. Eq. 1 generalises directly to F factors: for each candidate binding σ (a permutation of message positions), each position is penalised (KL to its unchanged distribution) for moving under interventions on the $F-1$ factors it is not bound to, and rewarded (hinge on a Jensen-Shannon margin, $\tau = 0.45$ throughout) for moving under the intervention on its own factor; the loss is the minimum over $\sigma \in \mathfrak{S}_F$. The pilot (§9) instantiates $F = 3$ ($3! = 6$ bindings). Nothing else in the training loop changes; the weight on the modularity term is 0.75 in all benches.

The gate recovery operator. Given a view $x_e = g_e(f) + \varepsilon$ of the factor vector f , the gate builds the intervention-response Jacobian J_e (column i is the view’s response to setting factor i from the baseline), recovers per-view factor estimates $z_e = J_e^+(x_e - x_e^0)$, and averages across views. The two ablations of §7.1 remove one signal each: controllability-only keeps a single view’s z_e (no purification); invariance-only keeps the cross-view-shared principal subspace of the raw views (no intervention alignment). The depth bench types objects by their $\text{do}()$ -response signature (nearest grounded exemplar in the averaged view space), which is the same controllable-and-invariant reading used non-parametrically.