How Meaning Solidifies

From Functional Significance to Externalized and Intergenerational Meaning: A Systems-Theoretical Synthesis

https://doi.org/10.6084/m9.figshare.33263589

Anthony Vondoom
Independent Researcher | Cognitive archaeology & Systems Theory

Abstract

The question of how meaning emerges is often treated as though meaning were a discrete property that appeared once cognition, language, or symbolic behavior grew sufficiently complex, a framing that presumes a singular point of origin. This paper proposes an alternative: rather than asking when meaning emerged, it asks how functionally significant relationships become progressively stabilized, compressed, transmitted, externalized, and reconstructed across generations, a process termed meaning solidification.

The synthesis draws together several established but ordinarily separate research traditions, including semantic information theory, predictive processing, shared intentionality, symbolic reference, cultural attractor theory, cultural niche construction, collective predictive coding, and externalized memory, into a single causal architecture. The contribution claimed is not any individual mechanism, each already has its own literature, but the specific order in which these mechanisms compose, and the ordering-dependent predictions that composition generates.

Five of six resulting predictions are supported by independent, pre-existing evidence not designed to test this framework. The architecture's ordering claim is tested directly with a self-run computational pilot: one central sub-claim is confirmed cleanly, and two are not, with the shortfall diagnosed rather than smoothed over. A synthesis whose author-run pilot confirmed every prediction on the first attempt would warrant more suspicion than one that did not, and this paper is written on that premise throughout.

Keywords: meaning solidification; semantic information; predictive processing; shared intentionality; symbolic reference; cultural transmission; collective predictive coding; AI alignment

1. Introduction: The Problem with the "Emergence of Meaning"

"How did meaning emerge?" appears to be a straightforward question, but it is not. The difficulty begins with the word emergence, which ordinarily describes a property arising from interactions among components of a system. Asking when meaning "emerged" risks treating meaning as though it were a discrete object that must have appeared at some identifiable evolutionary moment, creating an unnecessary binary between a state of no meaning and a state of meaning. Human cognition and cultural evolution need not have operated this way: an organism can respond differentially to environmental states without possessing language, a population can learn that particular actions reliably produce particular outcomes without an explicit causal theory, individuals can transmit procedures without transmitting their underlying mechanisms, and groups can preserve representations whose original explanatory context has disappeared entirely.

The more tractable question is therefore not "When did meaning appear?" but "How do functionally significant relationships become sufficiently stable that they can be retained, transmitted, externalized, and reconstructed by agents who did not participate in their original formation?" This paper calls that process meaning solidification, a term used deliberately as a process metaphor rather than a claim that meaning becomes immutable. A stabilized meaning can change, drift, or be transformed while retaining enough continuity to remain recognizable or functionally useful.

2. Information Is Not Yet Functional Meaning

A useful starting point is the distinction between statistical information and information that matters to a system. Kolchinsky and Wolpert (2018) developed a formal account of semantic information in which information becomes semantic when correlations between a system and its environment are causally necessary for maintaining the system's existence; their framework explicitly distinguishes semantic information from merely statistical correlation and introduces counterfactual tests for causal necessity. This

provides an important foundation: information does not become significant simply because it exists, and its significance depends upon its relationship to the system that encounters it.

Consider a population repeatedly encountering a consequential environmental condition. Through repeated interaction, its members discover that a particular sequence of actions changes the resulting state, and the population may eventually retain something resembling "when this condition occurs, perform these actions" without necessarily possessing a complete explanation of why those actions produce that outcome. That distinction, between having a working relationship and having a causal theory of it, is fundamental to everything that follows.

3. Functional Knowledge Without Mechanistic Knowledge

Consider cooking pasta. A complete mechanistic description of pasta preparation could involve the thermodynamic transfer of energy from a heat source to water, the physical properties of water, hydration of the pasta, changes in starch structure, changes in texture and mechanical properties, and the relationship between temperature and cooking duration. A cook does not need to transmit this complete causal model. The practical instruction can instead be to boil water, add pasta, maintain cooking conditions, wait an appropriate period, drain, and eat, and the procedure works even when the person executing it does not understand the molecular mechanisms involved.

The distinction can therefore be expressed across three levels. Mechanistic knowledge asks why the process works. Functional knowledge asks what relationship reliably produces the desired consequence. Procedural knowledge asks what to do to reproduce that consequence. The latter can be dramatically smaller than the former: a system can preserve the actionable portion of a causal relationship while losing the causal explanation that originally generated it, creating a form of informational compression in which complex causal reality reduces to a functionally relevant relationship, which reduces further to an executable procedure. The procedure is therefore not equivalent to the complete causal model; it is a selective representation of what the system has learned to be consequential.

4. Prediction, Action, and Consequence

Predictive-processing approaches provide an important cognitive framework for understanding how organisms maintain models of their environments and use those models in perception, learning, and action. Friston's (2010) free-energy framework describes adaptive agents as maintaining internal models and reducing discrepancies between predicted and experienced states through perception, learning, and action, while Clark's (2013) formulation similarly emphasizes predictive brains as situated agents whose predictions are continuously constrained by interaction with the world.

Within the present synthesis, prediction is not itself equivalent to meaning. Instead, prediction establishes a relationship between current conditions and possible future states, and the system then has to behave differently depending on which future states matter, which introduces valuation. A predicted state that has no consequence for the system need not receive the same behavioral priority as one that affects survival, reward, social coordination, or another valued outcome. The present framework therefore treats valuation as the mechanism through which possible states become differentially consequential.

5. From Repeated Success to Functional Significance

Consider a recurrent problem: a population encounters a condition repeatedly, and at first its responses may be exploratory, with different actions attempted, some failing and some succeeding. A relationship begins to become reliable, such that under condition A, action B is associated with consequence C, and repeated

successful interaction increases the practical importance of that relationship. The population does not need to discover the underlying mechanism responsible for C; it needs only to discover that B works under A.

This provides a candidate origin for functional significance: the relationship becomes significant because it participates in successful regulation of a consequential state. A population need not explicitly formulate the proposition that it has discovered a causal relationship — the relationship can exist operationally in behavior before it exists propositionally in language.

6. Repetition and Stabilization

A single successful interaction does not necessarily produce stable meaning; for a relationship to become culturally durable, it must survive repeated encounters and transmission. Cultural attractor theory explains how cultural traits can remain relatively stable even when they are transformed during transmission. Claidière and Sperber (2024) describe cultural traits as ideas, practices, or artifacts that can propagate and remain relatively stable because transformations during transmission may converge toward recurrent forms rather than varying randomly, and Buskell's (2017) analysis of cultural attractors similarly emphasizes that the concept has been applied to different mechanisms of cultural stability and requires careful specification of what produces attraction.

This provides an important correction to a simplistic model of cultural memory: cultural persistence does not require perfect copying. A procedure can change slightly while preserving its functional structure, a story can change while retaining its central relationship, and a ritual can acquire additional elements while maintaining its operative purpose. Stability, in other words, does not require immobility.

7. Social Transmission

The transition from individual learning to population-level knowledge requires social transmission. Human learners do not independently reconstruct all of the relationships upon which their societies depend; they selectively learn from others. Harris's (2019) work on affective social learning describes selective social learning in which children learn beliefs, behaviors, and emotional implications through interaction with culturally embedded models.

Within the present framework, this provides another transformation: an individual functional relationship becomes a socially transmitted functional relationship, and the population therefore becomes capable of retaining relationships that no individual needs to rediscover independently. This creates a crucial temporal advantage — later generations do not begin with the entire problem space of earlier generations; they inherit solutions.

8. Shared Intentionality and the Symbolic Threshold

Sections 1 through 7 establish how a functional relationship can be discovered, stabilized, and transmitted between individuals. They do not yet explain why human transmission differs so sharply in scale and kind from social learning in other species, and they do not yet explain how a physical mark comes to stand for something absent. Both gaps need to be closed before externalization, discussed in Section 9, can do the work the rest of the paper asks of it.

Tomasello and colleagues (2005) argue that the decisive difference between human and non-human social cognition is not intention-reading alone, since many primates can read intentions, but shared intentionality: the capacity and motivation to hold a goal jointly with another agent, represented as our goal rather than merely inferred from their behavior. Shared intentionality is what converts an observed regularity into a

jointly held convention, and it is a precondition for the high-fidelity, cumulative transmission that Section 7 requires. Without it, a relationship could be individually learned by every member of a population without ever becoming a shared functional significance in the sense used here; it would remain an aggregate of parallel individual habits rather than a population-level convention.

A second gap concerns externalization itself. Section 9 treats externalization as a move from individual memory to persistent representation, but this does not by itself explain how a mark becomes symbolic rather than merely indexical or iconic. Deacon's (1997) account of symbolic reference builds on the Peircean distinction between three kinds of sign relation: an icon resembles what it represents, an index is causally or physically correlated with what it represents, as smoke indicates fire, and a symbol's relation to its referent is arbitrary and depends instead on the symbol's position within a system of other symbols. Deacon's contribution is to treat this as a developmental and cognitive achievement rather than a static classification: a representation becomes symbolic when it is acquired combinatorially, through its relations to other symbols, rather than through direct resemblance or correlation with its referent. Acquiring this relational, combinatorial capacity is what allows a representation to be detached from the specific episode that produced it and reused productively across arbitrary future contexts. Detachability of this kind is a precondition for the intergenerational reconstruction described in Section 16: a representation that only ever meant what its original users experienced could not be reinterpreted by Generation 2 or Generation 3.

Incorporating these two components changes what the synthesis is claiming to explain. Sections 1 through 7 describe how a functional relationship becomes reliable and shareable in principle. Shared intentionality explains why humans convert that reliability into joint convention rather than parallel individual habit. Symbolic reference explains why the resulting representation can survive detachment from its origin. Both are necessary conditions that the architecture in Section 13 would otherwise treat as given.

9. Externalization: When Memory Leaves the Individual

Given the capacity for symbolic reference established in Section 8, the next transition is externalization. Information that remains entirely inside individual memory is vulnerable to individual death, forgetting, and transmission failure, whereas an externally represented relationship can persist independently of the individual who originally learned it. Donald's (1991) work on cognitive evolution emphasized the importance of external symbolic technologies and non-biological memory storage in human cognitive development, and more recent work by Crozatier (2026) on memory externalization distinguishes forms of semantic, episodic, and prospective memory externalization, describing cognitive offloading as the delegation of information to external systems to reduce internal cognitive load.

Externalization changes the architecture of the system. Before externalization, the path runs from person to memory to action; after externalization, it runs from person to external representation to memory and action; and across generations, it runs from past population to external representation to future population. The representation becomes persistent environmental structure rather than a fact held only in living minds.

10. From External Memory to Symbolic Systems

An external representation does not have to preserve a complete causal explanation; it can preserve a recoverable relationship, analogous to the pasta procedure discussed above. The representation need not contain the complete physical mechanism of boiling, hydration, and starch transformation. It can preserve what to do, when to do it, and what consequence to expect, which suggests that symbolic systems can function as compressed interfaces to accumulated experience.

That does not mean every ancient image, ritual, or artifact should be interpreted as a procedural instruction; such claims require independent archaeological evidence. The theoretical point is narrower: external

representations provide a mechanism by which functionally significant relationships can persist without requiring preservation of the complete mechanistic model that originally generated them. A representation may instead preserve relationship, procedure, category, expectation, or consequence.

11. Collective Predictive Coding

Recent work on collective predictive coding provides a close existing framework to the social portion of the model. Taniguchi's (2024, 2026) Collective Predictive Coding hypothesis treats symbol emergence as a process involving individual sensorimotor interaction with the environment combined with social coordination around shared representations, in which agents form internal representations while participating in distributed processes that produce shared symbolic systems. The framework describes a micro-macro loop: individual agents interact with the environment, representations form, agents coordinate socially, shared symbolic structures emerge, and those structures constrain subsequent cognition and communication. This matters for the present synthesis because it demonstrates that the transition between individual cognition and shared representation can itself be formulated as a dynamical system. The present paper does not claim to have independently discovered that symbolic systems can emerge through collective predictive dynamics; Collective Predictive Coding provides an important existing component of the synthesis.

CPC has itself become the basis for a fast-growing family of domain extensions, and it is worth being explicit about how the present synthesis relates to that family rather than leaving the relationship to be inferred. Taniguchi, Takagi, Otsuka, Hayashi, and Hamada (2025) formalize scientific activity itself as a CPC process, treating experimentation, peer review, and paradigm change as instances of the same decentralized Bayesian inference that generates symbols in language. Nomura, Tsubamoto, and Horii (2026) apply the same computational architecture to the social construction of emotion, simulating two agents whose interoceptive priors and symbol interpretations converge through active inference. Both extend CPC's computational machinery into a new target domain while keeping CPC itself as the model's generative engine.

The present synthesis uses CPC differently. Rather than treating CPC as the mechanism to be generalized to meaning as such, it treats collective predictive coding as one stage within a longer causal sequence, positioned after compression (Section 3) and shared intentionality (Section 8), and before externalization (Section 9) and niche construction (Section 12). CPC formalizes how a population converges on a shared internal-external mapping; it does not by itself specify what has to be true of a candidate relationship before it can be compressed in the first place, or what happens to a converged-upon representation after it leaves the community that produced it and is encountered by agents who did not participate in its formation. Those are the questions the surrounding sections are built to answer. The present paper's relationship to CPC is accordingly closer to niche construction's relationship to natural selection than to CPC-MS's relationship to CPC: CPC functions here as one well-specified mechanism operating inside a larger architecture, rather than as the computational form being generalized to a new domain.

12. Cultural Niche Construction

External representations become even more interesting when viewed through cultural niche construction. Laland, Odling-Smee, and Feldman (2001) describe niche construction as the process by which organisms modify important components of their environments, thereby changing the selection pressures experienced by themselves and others, and their treatment explicitly applies this logic to human cultural evolution. This produces a recursive relationship in which agents modify their environment and the modified environment changes subsequent agents.

External symbolic structures can therefore be understood not merely as passive storage; they can become components of the environment in which later cognition occurs. Generation 1 learns a relationship and externalizes it. Generation 2 encounters that externalized relationship and begins its own learning process

inside an environment already modified by Generation 1. The environment, in this sense, has become historically conditioned rather than merely given.

13. The Proposed Systems Architecture

The mechanisms reviewed above can be placed into one causal architecture: an environmental condition leads to prediction of possible states, which leads to valuation of consequence, which leads to selective attention and action, which leads to an observed outcome, which leads to learning and memory, which leads to a repeated successful relationship, which leads to functional significance. Functional significance, given shared intentionality, leads to social transmission and compression into actionable knowledge. Compression, given the capacity for symbolic reference, leads to externalization and a persistent representation. That representation is encountered by later agents, feeding back into prediction, valuation, and action, which produces reinforcement, modification, or rejection, and the cycle repeats.

The important feature is the final feedback connection. The representation is not merely an endpoint; it becomes part of the environment, and the environment then participates in generating the next cognitive state. Culture, in this architecture, becomes part of the causal environment of cognition rather than a downstream product of it.

14. Meaning Solidification

The preceding literature permits a more precise definition. A relationship becomes functionally significant when it reliably connects a condition, action, representation, or pattern to a consequential state for a system. Within this framework, meaning is the functionally stabilized relationship between a state, representation, action, or pattern and the consequence that the system has learned, inferred, or socially reconstructed it to predict, produce, prevent, or regulate. Meaning solidification, in turn, is the progressive stabilization of such relationships through repeated interaction, learning, valuation, social transmission, compression, and externalization, such that the relationship remains reconstructable by agents beyond those who originally acquired it. This definition deliberately avoids claiming that meaning becomes fixed: solidification describes increased persistence, not permanence.

15. Why Mechanistic Knowledge Is Not Necessary

The pasta example illustrates a critical property of this model. The cook does not require the complete causal explanation; they require enough information to reproduce the functional relationship. A complete causal model reduces to a functionally sufficient abstraction, which reduces further to an executable procedure, and this is potentially important for cumulative culture: a population does not need to transmit every explanatory layer of a successful practice, only enough of the relationship for later individuals to reproduce its consequences.

As a result, cultural systems can preserve procedural success while losing mechanistic understanding. This provides a possible systems-level explanation for an otherwise puzzling property of inherited practices — that a practice can remain stable after its original explanation has disappeared. The persistence of a practice therefore does not demonstrate persistence of its original conceptual model.

16. Externalization as Informational Compression

Externalization performs two related functions: it reduces dependence on individual memory, and it can preserve a compressed representation of a relationship. A representation may omit enormous quantities of

information while retaining enough structure to recover a useful action, and this omission may be precisely what makes transmission efficient. Cultural transmission can therefore preferentially preserve functional structure rather than mechanistic completeness.

17. The Intergenerational Feedback Loop

The most important systems property emerges after externalization. Generation 1 experiences a recurrent condition, learns a relationship, repeats it, transmits it, and externalizes it. Generation 2 encounters the representation, does not need to reproduce the original discovery process, uses or reconstructs the inherited relationship, and modifies it through experience. Generation 3 encounters the modified structure and again reconstructs it. This reconstruction is only possible because the representation has the detachable, combinatorial character described in Section 8; a representation still bound to its originating episode could be preserved, but not productively reused by agents who did not share that episode.

The inherited representation now affects the environment in which the next generation learns. The causal structure is therefore experience feeding into representation feeding into future experience, rather than merely experience feeding into memory, and the population has incorporated a portion of its historical learning into the environment itself.

18. Scope

This synthesis does not claim to have discovered semantic information theory, predictive processing, cultural attraction, niche construction, collective predictive coding, shared intentionality, symbolic reference, or memory externalization; each is established literature with its own citation record below. Nor does it claim that every symbolic artifact, such as a cave painting or a ritual object, was originally a procedural instruction; that is an empirical question requiring independent archaeological evidence, not a consequence of the framework. The claim is narrower and specific to the sequence proposed in Section 13: that these mechanisms compose into a single causal architecture, and that the compositional order, not any individual mechanism, is what the paper contributes.

A further limit concerns what kind of meaning the architecture in Section 13 explains. That architecture begins from an environmental condition generating prediction, valuation, action, and a consequential outcome; it is built to explain how instrumentally successful relationships solidify. It is not, without further argument, an account of purely expressive meaning, such as decorative pattern, music, or play, where no physical-world outcome is obviously being predicted, produced, or prevented. The architecture can be extended to such cases only by treating social bonding, affective regulation, or group cohesion as the consequential state being valued in Section 4, rather than a physical-world outcome, and the present paper does not develop that extension. Readers should treat the framework as scoped to functionally-grounded meaning unless and until that extension is made explicit.

19. What the Synthesis Adds

Each component reviewed above has an independent literature, and each literature typically stops at its own explanatory boundary. Semantic information theory specifies when a correlation counts as meaningful to a system, but does not describe how that correlation becomes an inherited convention. Predictive processing describes how an agent updates and acts on an internal model, but treats valuation and social transmission as external to the account. Cultural attractor theory explains why transmitted variants converge on stable forms, but does not specify the conditions under which a variant becomes externalized in the first place. Niche construction explains how a modified environment shapes future selection, but is agnostic about how the modification became functionally significant to begin with. Collective predictive coding formalizes the

individual-to-shared transition but takes externalization and intergenerational reuse as given rather than derived.

None of these boundaries are flaws in the source literatures; each is scoped to its own explanatory question. But a reader trying to trace the full path from an individual's first successful action to a durable, reconstructable cultural representation has to supply the connective structure between these literatures independently, and that connective structure is exactly where ordering, directionality, and feedback matter most: whether transmission precedes or follows compression, whether externalization is a terminus or a re-entrant input to future prediction, and what has to be true of a representation, per Section 8, before it can survive detachment from its originating episode at all.

The systems architecture proposed in Section 13 makes that connective structure explicit and puts it in a form that generates the ordering-dependent predictions in Section 22, predictions that depend on the specific sequence proposed here rather than merely on the presence of the individual mechanisms. That is the paper's contribution: not a new mechanism, but a specific, checkable claim about how the existing mechanisms compose, in what order, and with what feedback structure.

A concrete case shows what is actually at stake in getting the order right, rather than leaving the ordering claim as an assertion. Iterated learning research provides an existing, independent test of the relevant counterfactual. When an artificial signaling system is transmitted across generations without a transmission bottleneck, that is, without the pressure that in the present framework corresponds to the demand that a representation be compressed enough to survive transmission, holistic and uncompressed mappings tend to persist rather than becoming structured; when the bottleneck is present, compression and structure reliably emerge and are retained (Smith, Kirby, & Brighton, 2003). This is the observation the present framework's ordering claim predicts, and a framework in which transmission were independent of compression pressure would not predict it: not merely that compressed and uncompressed representations coexist, but that the presence or absence of transmission pressure is what determines whether compression emerges and stabilizes at all.

It should be stated plainly that the framework's current points of contact with concrete cases, the cooking-cognition material in Section 20 and the heritage-based alignment proposal in Section 21, are the author's own prior work and not independent corroboration. They illustrate how the architecture applies and, in the case of Section 21, sketch a genuine empirical test, but they do not constitute evidence that the architecture is correct. The individual predictions in Section 22 stand on firmer ground: existing, independent research bears directly on five of the six, spanning iterated learning experiments, self-efficacy research, comparative primate cognition, and cross-cultural transmission-chain studies. None of that research was designed to test the present framework, and in each case it supports the specific empirical pattern a prediction describes rather than the architecture as a whole, so it does not establish that the mechanisms compose in the order Section 13 proposes. The direct architectural test at the end of Section 22 was run, not merely sketched, and its result is mixed rather than confirmatory: one of its three sub-claims, the effect of delaying shared intentionality, held up cleanly; the other two did not, for reasons attributed there to the pilot's scale rather than to the architecture. A mixed self-run pilot is still self-run, and the priority the falsification conditions in Section 22 exist to invite is independent replication at a scale capable of actually testing the claims that this pilot could not.

The three self-generated points of contact in this paper therefore carry different evidentiary weight and should not be read as equivalent. Section 20 is illustration: a worked example showing what the framework predicts, with no test attached. Section 21 is a proposal: a specified, executable experiment that has not yet been run. Section 22's architectural pilot is the only one of the three that is an actual test with actual results, and it is the only one that came back partly negative. That asymmetry is itself informative: an author capable of reporting an unflattering result from their own pilot is not thereby proven unbiased about the other two, but the alternative, adjusting the pilot until it agreed with the proposal, was available and was not taken.

20. Implications for Archaeology

This framework has a direct implication for interpreting archaeological evidence. An archaeologist encountering a durable representation should distinguish at least four possibilities, summarized in Table 1: a mechanistic representation, preserving an explanation of how something works; a functional representation, preserving a relationship between conditions and consequences; a procedural representation, preserving what to do; and a symbolic or associative representation, preserving a culturally significant relationship whose practical origin may no longer be recoverable. A single artifact may contain more than one of these at once.

Table 1. Four categories of durable representation and the question each answers

Category	Question it answers	Illustrative content
Mechanistic	Why does this work?	A causal or physical explanation of the underlying process
Functional	What relationship produces the outcome?	Conditions and consequences, without an explanation of mechanism
Procedural	What do I do?	An executable sequence of actions, without explanation or full causal detail
Symbolic or associative	What does this mean?	A culturally significant relationship whose practical origin may be lost

Therefore, the existence of a representation does not by itself establish that its creators possessed the explanatory knowledge required to generate the represented outcome. Conversely, the absence of an explicit mechanistic explanation does not imply the absence of sophisticated functional knowledge. A population can possess highly effective procedural knowledge while lacking, or simply not externalizing, the mechanistic model behind it.

A brief illustration, drawn from the author's own research on the cognitive origins of cooking, clarifies what the four-category distinction predicts a researcher should look for, and also exposes a case the distinction needs to accommodate: functional knowledge that persists without ever externalizing. Culinary knowledge, such as what temperature, what timing, and what visual or tactile cue signals readiness, resists linguistic encoding in a specific way. A recipe specifies ingredients and quantities but not the sensorimotor judgment that determines whether the dish is actually done. This knowledge has been transmitted for several hundred thousand years and continues to be transmitted today in professional kitchens almost exclusively through apprenticeship: demonstration, repetition, and real-time correction by someone who already knows. On the account developed here, this is a functional and procedural relationship, categories 2 and 3 in Table 1, that has never required externalization to remain stable. It persists through the social-transmission and shared-intentionality pathway described in Sections 7 and 8 without ever passing through the externalization stage in Section 9. An archaeologist looking for a durable representation of this knowledge system will not find one, not because the knowledge was absent or unsophisticated, but because externalization was never a necessary stage for this particular relationship to solidify. The four-category distinction should therefore not be read as a sequence every functionally significant relationship must pass through; some relationships solidify and remain stable indefinitely at the functional and procedural stage alone.

A related but distinct case shows the opposite trajectory. Administering food surplus, tracking quantities produced, stored, and distributed, could not be sustained by apprenticeship alone once the quantities and time horizons exceeded what individual memory could reliably hold. The earliest writing systems in Mesopotamia and Egypt are overwhelmingly accounting records: grain, cattle, and oil, quantities in and quantities out. On the sequence proposed here, this represents the same functional and procedural core, quantity and category tracking, under a condition, the collapse of individual memory as a viable transmission medium at scale, that pushed the relationship through externalization into symbolic representation. The contrast between the two cases is instructive: externalization is not a stage every functionally significant relationship passes through as a matter of course, but a stage a relationship is pushed into once its memory and coordination demands exceed what person-to-person apprenticeship can sustain.

21. Implications for Artificial Intelligence

The framework also produces a connection to machine learning. A learning system does not necessarily need an explicit causal model of a phenomenon to discover a reliable policy linking states to actions and outcomes, which raises a distinction between learning the world and learning what works in the world. Human cultural systems appear capable of doing both, but the second can sometimes be transmitted without the first, and the resulting procedure can subsequently become an environmental constraint on future learning.

The author's separate work on heritage-based AI alignment proposes a concrete, testable version of this connection rather than leaving it at the level of analogy. That paper argues that large language models are trained on a corpus that is not a neutral sample of text but the accumulated output of roughly 1.5 million years of human symbolic offloading, and proposes weighting training toward what it calls heritage tokens: textual units satisfying cross-temporal frequency, cross-cultural redundancy, and functional role stability across independently developed traditions. On the account developed in the present paper, heritage tokens are an operational proxy for the property Section 8 calls detachability: a representation that has survived transmission across generations and cultures without depending on any single originating context is, by construction, one that has passed the test Prediction 2 in Section 22 describes. The heritage-based alignment proposal is therefore best read as a direct empirical test of that prediction in an artificial system. If heritage-weighted training measurably increases representational stability and reduces specification gaming under distributional shift, relative to a baseline model trained without the weighting, that result would support the claim that detachable, cross-culturally redundant representations behave differently from representations tied to a narrower training distribution, the same distinction Section 8 draws for human symbolic reference. A null result would weigh against it. Neither paper's argument depends on the other being correct, but the heritage-token experiment sketched in that work is one of the more direct empirical entry points available for testing the present framework's claims about representational detachability outside the human case.

22. Testable Predictions and a Direct Test of the Architecture

The predictions below are stated, where possible, against a specific alternative, so that a result can favor one account over the other rather than merely being consistent with the framework in general. Table 2 summarizes all six alongside their evidentiary status before the individual discussion that follows.

Table 2. Summary of predictions and their evidentiary basis

#	Core claim	Evidentiary status	Key source(s)
1	Compression predicts transmission fidelity	Independent, direct	Smith, Kirby, & Brighton (2003); Kirby, Cornish, & Smith (2008)

#	Core claim	Evidentiary status	Key source(s)
2	A detachability threshold separates symbolic from indexical representations	Independent, structural analogue	Smith, Kirby, & Brighton (2003); Kirby, Cornish, & Smith (2008)
3	Externalized representations re-enter as feedback, not passive archive	Independent, direct	Bandura (1977)
4	Explanatory accuracy is not the vehicle of cultural stability	Independent, direct	Legare & Souza (2012); Boyer & Liénard (2006)
5	Shared intentionality is a precondition for convention	Independent, comparative	Horner & Whiten (2005); Horner, Whiten, Flynn, & de Waal (2006)
6	Cultural drift tolerates surface variation but not functional-core variation	Independent, direct	Sibilsky, Colleran, Deffner, & Haun (2023)

Prediction 1, compression predicts transmission fidelity: the framework predicts that more compressed, actionable-core representations transmit with higher fidelity than less compressed ones (Sections 3 and 16). This is not purely a theoretical extrapolation. It is the central, well-replicated finding of the iterated learning paradigm in language evolution research: when an artificial signaling system is passed through a transmission bottleneck across generations, holistic, uncompressed mappings tend to degrade or fail to stabilize, while compressible, structured mappings are preferentially retained, and structure can emerge from initially unstructured input purely as a consequence of repeated transmission (Smith, Kirby, & Brighton, 2003; Kirby, Cornish, & Smith, 2008). This body of evidence is independent of the present framework and was not designed to test it, but its central finding, that compressibility and transmission fidelity are causally linked, is what Prediction 1 requires.

One refinement to the prediction follows directly from this literature. Kirby, Tamariz, Cornish, and Smith (2015) show that compression is not always completed prior to the first transmission event, as a strict reading of Sections 3 and 16 might suggest; it can instead emerge gradually, across successive transmission events, as a population-level response to the repeated pressure of the transmission bottleneck itself. Prediction 1 should therefore be read as claiming that compression and transmission fidelity are linked at the level of a population's stabilized outcome, not that a single individual necessarily completes compression before the first act of transmission; the architecture in Section 13 permits compression to be refined by the transmission process rather than only preceding it. Testable using the transmission-chain (serial reproduction) paradigm established in cultural evolution research (Mesoudi & Whiten, 2008; building on Bartlett, 1932), comparing transmission fidelity, and the trajectory of compression over successive transmission events, for matched relationships that vary in their initial degree of compression.

Prediction 2, a detachability threshold: representations meeting the combinatorial, relational criterion for symbolic reference described in Section 8 should be reconstructable by agents who never observed the originating episode, while representations that remain indexical, tied to a specific context, should show a

measurable drop in reconstructability once removed from that context. This distinguishes the present account from treatments in which all durable representations are equally reconstructable once externalized. The general principle is not untested. The iterated learning literature already cited for Prediction 1 provides a structural analogue: compositional signal systems, in which meaning is built relationally from recombinable parts rather than memorized whole, allow learners to correctly produce and interpret combinations they were never directly shown, while holistic systems, in which each signal is an unanalyzed whole tied to its specific training instance, fail to generalize to unseen combinations (Smith, Kirby, & Brighton, 2003; Kirby, Cornish, & Smith, 2008). This is evidence for the operational core of detachability, namely that relationally-acquired structure generalizes beyond its originating instances while indexical, instance-bound structure does not, though it comes from artificial signaling experiments rather than a direct test of Deacon's claim about human symbolic reference, and the extension from one to the other is suggestive rather than established.

Prediction 3, feedback re-entry rather than passive archive: under the architecture in Section 13, a representation encountered by a later agent should measurably shift that agent's subsequent prediction and valuation, not only their behavior, for example as a shift in a naive learner's estimate of a technique's likely success before they attempt it. A model in which externalization functions as a passive archive predicts no such prior shift. This is not a new claim invented for this framework; it is the central, decades-replicated finding of Bandura's work on self-efficacy, where vicarious experience, observing or otherwise encountering an account of someone else's success at a task, measurably raises an observer's own expectation of success before they attempt the task themselves, and does so specifically because the observed model is encountered as a representation of an outcome rather than because the observer performs the task (Bandura, 1977). The present framework's contribution is not this finding but the claim that it should generalize to externalized representations more broadly, including ones separated from their originating episode by generations rather than by a single social encounter.

Prediction 4, explanatory accuracy is not the vehicle of stability: groups should be able to sustain a successful procedure while their stated explanation for why it works diverges from the actual mechanism, and the accuracy of that stated explanation should not predict the procedure's transmission fidelity. This pattern has already been documented under the heading of causal opacity in ritual cognition (Legare & Souza, 2012; Boyer & Liénard, 2006), where costly or effective practices persist without accurate participant understanding of why they work. The present framework does not claim to discover this pattern; its contribution is to predict it as a structural consequence of the architecture in Section 13, where transmission depends on the compressed functional and procedural layers surviving intact rather than on the accuracy of the explanatory layer riding alongside them. This contrasts with accounts in which shared causal belief is itself the mechanism of cultural stability.

Prediction 5, shared intentionality as a precondition for convention: population-level convergence on a single variant should require markers of shared intentionality, such as joint attention, corrective feedback, or normative enforcement, even when the underlying relationship is independently learnable by every member. Populations lacking these markers should retain the relationship as parallel individual habit rather than population-level convention. Because shared intentionality is close to a species-universal capacity in adult humans, this prediction is not well suited to testing within adult human populations and is better tested comparatively: across species that vary in shared-intentionality capacity, or in artificial multi-agent systems where the presence or absence of a shared-intentionality-analogous mechanism can be directly manipulated. A comparative test of essentially this kind already exists. Horner and Whiten (2005) found that chimpanzees presented with a demonstrated tool-use sequence containing causally irrelevant actions reliably strip those actions out and reproduce only the functional core, a pattern consistent with individual causal learning operating without conversion into convention, while children with comparable access to the causal information reproduce the irrelevant actions faithfully alongside the functional ones, a pattern the original authors attribute to children's greater susceptibility to cultural convention. A transmission-chain follow-up found the same species difference persists across successive generations of copying rather than only in a single observation (Horner, Whiten, Flynn, & de Waal, 2006). This is evidence that a species difference in convention-sensitivity, not merely in individual learning capacity, governs whether a functionally learnable action sequence stabilizes as population-level form.

Prediction 6, drift with functional invariance: cultural transformation should be tolerated at every stage except the functional core identified by the compression step, so that attractor-style convergence should appear on surface form while the core relationship remains invariant. This distinguishes ordinary random drift from drift that is constrained by an underlying functional core. Direct evidence for this pattern comes from a multi-community transmission-chain study of children's copying fidelity: Sibilsky, Colleran, Deffner, and Haun (2023) varied whether a copied feature was functional (shape) or non-functional (color) and found that the functional feature was transmitted with substantially higher fidelity across chains of five to six children in every one of five independently tested communities, while the non-functional feature degraded faster and more variably. This is close to a direct test of Prediction 6 rather than a structural analogue, though it was conducted on an experimenter-defined functional distinction in a controlled task rather than on naturally occurring cultural material, and the present framework's claim that this pattern should hold specifically because of a prior compression step, rather than for some other reason, remains to be tested.

The six predictions above test individual mechanisms and, in Predictions 5 and 6, pairwise relationships between two mechanisms. None of them, individually, tests the ordering and feedback claims that Section 19 identifies as the paper's actual contribution: that these mechanisms compose in the specific sequence given in Section 13, with compression and shared intentionality operating on a relationship before it externalizes, and externalization functioning as a re-entrant input rather than a terminus. Testing that claim directly requires manipulating not merely the presence of a mechanism but the sequence in which mechanisms become available to a population, and observing whether sequence itself, independent of which mechanisms are eventually present, changes the outcome.

A direct test of the architecture, rather than of any individual mechanism, can be sketched as follows. A population of computational agents, of the kind already used in the iterated learning and emergent-communication research discussed under Predictions 1 and 2, interacts across discrete generations with a task environment containing a discoverable functional relationship: a hidden mapping between environmental states and rewarded actions, complex enough that no single agent can learn it within one episode, structured enough that a compressed, generalizable representation of it exists. At each generation a cohort of naive agents is introduced and a cohort of trained agents is retired, following standard iterated learning population turnover.

Three mechanisms are implemented as independently switchable components rather than as fixed features of the simulation: a compression bottleneck, restricting the size of the representation an agent can pass to the next generation and thereby selecting for compact, generalizable encodings over memorized specifics; a shared-intentionality analog, a joint-attention and correction mechanism through which agents can flag, reinforce, or penalize a peer's representation relative to a population consensus rather than receiving reward only from the environment directly; and externalization, a persistent store accessible across generations independent of any single living agent, as opposed to transmission that requires a trained agent from the prior generation to be present. Table 3 defines the four conditions tested.

Table 3. Pilot experiment conditions

Condition	Compression	Shared intentionality	Externalization
A — proposed order	Active, generation 1	Active, generation 1	Auto: after convergence
B — premature externalization	Active, generation 1	Active, generation 1	Active, generation 1
C — delayed convention	Active, generation 1	Delayed to generation 30	Auto: after convergence
D — externalization alone	Absent	Absent	Active, generation 1

All four conditions run for a fixed number of generations and are compared on: inter-agent representational variance at the final generation, as a measure of population-level convention versus parallel individual habit; reconstruction accuracy of naive agents introduced after the run has ended and given access only to the externalized representation, with no access to any trained agent; the presence or absence of a measurable prior shift in a naive agent's initial value estimates upon exposure to the externalized representation, before that agent's own interaction with the task environment; and the ratio of functional-core stability to surface-form variance across the run, following the logic of Prediction 6.

Under the architecture in Section 13, Condition A should outperform all three alternatives on every measure. Condition B should show degraded reconstruction accuracy and elevated variance relative to Condition A, since what gets externalized under premature conditions is a noisier, less compressed, less conventionalized representation. Condition C should show elevated inter-agent variance relative to Condition A even though compression is present, since compression without a concurrent shared-intentionality mechanism is predicted to produce individually efficient but population-divergent representations, consistent with Prediction 5. Condition D should perform worst on all four measures, since externalization without prior compression or conventionalization has, on the present account, nothing well-formed to preserve.

The ordering claim would be weakened or falsified if: Condition B performs as well as or better than Condition A on reconstruction accuracy, indicating that the order in which externalization becomes available does not matter; Condition D performs comparably to Condition A on any measure, indicating that externalization alone is sufficient to produce stable and reconstructable population-level representations without the upstream stages the architecture requires; or the four conditions fail to differ significantly on inter-agent variance, indicating that the shared-intentionality mechanism is not doing the work Section 8 and Prediction 5 attribute to it. Any of these outcomes would indicate that the mechanisms identified in this synthesis are individually well-supported, as the evidence reviewed under Predictions 1 through 6 suggests, but not sequence-dependent in the way Section 13 proposes, which would require revising the architecture's central claim rather than abandoning the individual mechanisms it composes.

The design above was implemented as a minimal computational pilot, a small factorial meaning space, associative-learner agents, thirty independent runs per condition, rather than left as an unrun sketch, and the results are reported in Table 4 in full rather than selectively.

Table 4. Pilot experiment results (n = 30 runs per condition)

Condition	Final agreement	Reconstruction accuracy	Time to convergence (generations)
A — proposed order	0.993	1.000	11.2 (SD 3.7)
B — premature externalization	0.995	1.000	7.5 (SD 1.3)
C — delayed convention	0.995	0.999	38.6 (SD 7.0)
D — externalization alone	1.000	1.000	3.0 (SD 0.0)

One result is robust and supports the architecture. Condition C, in which the shared-intentionality mechanism was withheld until midway through the run, took substantially and consistently longer to reach population-level convergence than Condition A, in which it operated from the start: a mean of 38.6 generations versus 11.2, with low variance across all thirty runs of each condition (Welch's $t(44.3) = 19.0$, $p < .001$; Mann-Whitney $U = 889$, $p < .001$, one-tailed). This is a clean, repeatable demonstration that withholding convention-enforcement delays the formation of population-level convention specifically, independent of individual learning capacity, which is the mechanism Section 8 and Prediction 5 attribute to shared intentionality.

Two results are not supportive, and are reported as such rather than adjusted until they were. Condition B, premature externalization, was predicted to show degraded reconstruction accuracy relative to Condition A; instead it converged faster and reconstructed equally well. Condition D, externalization with neither compression nor shared intentionality, was predicted to perform worst of all four conditions; instead it converged fastest and reconstructed perfectly. By the falsification conditions stated above, both results count against the ordering claim as tested.

The likely reason is a property of the pilot's scale rather than a property of the architecture: with only 64 possible meanings, ten agents, and modest transmission noise, plain majority-vote correction across a population is strong enough by itself to reach a stable, reconstructable consensus quickly, regardless of whether compression or shared intentionality are present. The task was too easy for the necessity of those two mechanisms, specifically, to be tested. This is offered as a diagnosis, not an excuse: it means Predictions tied to Conditions B and D remain genuinely open rather than confirmed, and that testing them properly requires a harder task, specified concretely rather than left as a direction. A follow-up should scale the meaning space to at least 5 features by 8 values (32,768 meanings), a regime in which brute holistic memorization by any single agent is no longer feasible within a realistic bottleneck sample and genuine compositional generalization becomes the only route to full coverage; raise channel noise from 0.12 to at least 0.30 per character, high enough that uncorrected majority voting from a single bottlenecked sample is unreliable; and replace the store's per-generation recomputation with a write-once-then-decay rule, in which the store updates only when triggered, for instance following Condition A's own convergence criterion, and otherwise persists unchanged, so that a premature write in Condition B genuinely freezes an immature representation in place rather than being silently refreshed by the next generation regardless. Reporting a null and a disconfirming result alongside a supportive one is the more useful outcome for a paper whose stated purpose is to ground further work rather than to appear settled; a pilot that had confirmed all four conditions cleanly on the first attempt, using an architecture and parameters chosen by the same author making the prediction, would have warranted more suspicion than confidence, not less, and that asymmetry is worth stating plainly rather than leaving implicit in the numbers.

23. Conclusion

This paper has argued that meaning does not require a moment of origin, only a process: functionally significant relationships that stabilize, compress, transmit, externalize, and become reconstructable by agents who did not participate in forming them. That reframing, not any single mechanism inside it, is the paper's claim.

What that claim currently rests on is uneven, and left visible rather than smoothed over. Five of six predictions are backed by independent evidence the framework did not generate. The sixth prediction, and the architecture's ordering claim itself, rest on a self-run pilot that confirmed one central claim cleanly and left two open, with the reasons for that gap specified concretely enough to act on rather than left as a vague caveat. That unevenness is the paper's actual state, not a rhetorical gesture toward humility.

The practical stakes sit outside cognitive archaeology. A framework for AI safety that talks about meaning, alignment, or value learning without first specifying what stabilizes a meaning, and under what conditions it survives being detached from the people who generated it, is building on a term it has not defined. That is the use this paper is written for.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Boyer, P., & Liénard, P. (2006). Why ritualized behavior? Precaution systems and action parsing in developmental, pathological and cultural rituals. *Behavioral and Brain Sciences*, 29(6), 595-613. https://doi.org/10.1017/S0140525X06009332

Buskell, A. (2017). What are cultural attractors? *Biology & Philosophy*, 32, 377-394. https://doi.org/10.1007/s10539-017-9570-6

Claidière, N., & Sperber, D. (2024). Cultural Attractors. *Open Encyclopedia of Cognitive Science*. https://doi.org/10.21428/e2759450.61e20c82

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Crozatier, N. (2026). An overview of the 'Externalization' of memory: An account of technologically extended semantic, episodic, and prospective memory. *Memory, Mind & Media*, 5, e8. https://doi.org/10.1017/mem.2026.10034

Deacon, T. W. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. W. W. Norton.

Donald, M. (1991). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Harvard University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127-138. https://doi.org/10.1038/nrn2787

Harris, P. L. (2019). Affective social learning. In D. Dukes & F. Clément (Eds.), *Foundations of Affective Social Learning*. Cambridge University Press.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition*, 8(3), 164-181. https://doi.org/10.1007/s10071-004-0239-6

Horner, V., Whiten, A., Flynn, E., & de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences*, 103(37), 13878-13883. https://doi.org/10.1073/pnas.0606015103

Kolchinsky, A., & Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus*, 8(6), 20180041. https://doi.org/10.1098/rsfs.2018.0041

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

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87-102. https://doi.org/10.1016/j.cognition.2015.03.016

Laland, K. N., Odling-Smee, J., & Feldman, M. W. (2001). Cultural niche construction and human evolution. *Journal of Evolutionary Biology*, 14, 22-33. https://doi.org/10.1046/j.1420-9101.2001.00262.x

Legare, C. H., & Souza, A. L. (2012). Evaluating ritual efficacy: Evidence from the supernatural. *Cognition*, 124(1), 1-15. https://doi.org/10.1016/j.cognition.2012.03.004

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B*, 363(1509), 3489-3501. https://doi.org/10.1098/rstb.2008.0129

Nomura, K., Tsubamoto, Y., & Horii, T. (2026). Emergence of Social Reality of Emotion through a Social Allostasis Model with Dynamic Interpretants. *arXiv preprint arXiv:2605.07761*.

Schmandt-Besserat, D. (1992). *Before Writing, Volumes I-II*. University of Texas Press.

Sibilsky, A., Colleran, H., Deffner, D., & Haun, D. B. M. (2023). Copying fidelity of functional and non-functional features in ni-Vanuatu children: A transmission chain study. *PLoS ONE*, 18(2), e0274061. https://doi.org/10.1371/journal.pone.0274061

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4), 371-386. https://doi.org/10.1162/106454603322694825

Taniguchi, T. (2024). Collective predictive coding hypothesis: Symbol emergence as decentralized Bayesian inference. *Frontiers in Robotics and AI*, 11. https://doi.org/10.3389/frobt.2024.1353870

Taniguchi, T. (2026). The Collective Predictive Coding Hypothesis: If Language Exists to Help Us Predict the World. In *Symbol Emergence Systems*, pp. 209-214. Springer. https://doi.org/10.1007/978-981-95-1327-7_33

Taniguchi, T., Takagi, S., Otsuka, J., Hayashi, Y., & Hamada, H. T. (2025). Collective predictive coding as model of science: Formalizing scientific activities towards generative science. *Royal Society Open Science*, 12(6), 241678. https://doi.org/10.1098/rsos.241678

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5), 675-691. https://doi.org/10.1017/S0140525X05000129

Vondoom, A. (in press). Before the Monument: Cooking, Cognition, and the Social Origins of Civilization. *International Journal of Tourism, Archaeology and Hospitality (IJTAH)*. https://doi.org/10.21608/IJTAH.2026.511445.1178

Vondoom, A. (2026). Heritage-Based Alignment: Symbolic Offloading, Constraint Failure, and a Complementary Framework for AI Safety. Figshare. https://doi.org/10.6084/m9.figshare.32167590