Deep Well Labs

INDEPENDENT RESEARCH + ENGINEERING

Thought-Science Research Map

What science currently knows, disputes, and does not know about human thought.

Public edition. This document states the company's position and the reasoning behind it. It is a working document, not a specification, and nothing in it is a claim that a described capability is implemented. Where Deep Well Labs describes what could become possible, that is research; where Origo describes what it does, that is product. The distinction is the subject of the Boundary Charter.

AUGUST 2026

DEEP WELL LABS, INC.

DEEPWELLLABS.COM

What this document is

A structured survey of what science currently knows, disputes, and does not know about human thought. It exists so that Deep Well Labs research and future Origo product reasoning can be grounded in the actual state of the evidence rather than in the popular summary of it.

This is a research map. It is not a product plan, an architecture decision, or a roadmap. It proposes no features. Where it touches Origo at all, it does so in a single clearly bounded section that raises questions rather than answering them.

The governing constraint is the Deep Well Labs and Origo boundary charter: research may explore what could become possible, but nothing in this document authorizes a product claim.

The central finding, stated first

There is no scientific consensus on what a single thought is.

"Thought" is not a natural kind in the way that "blood glucose" is a natural kind. It is a folk-psychological umbrella term that different sciences operationalize differently, and the operationalizations do not reduce to one another. Cognitive psychology studies it as information processing over representations. Neuroscience studies the neural population activity that accompanies task performance. Linguistics studies the structures available to express it. Philosophy of mind studies what makes it about anything at all.

These are not four views of one agreed object. They are four research programs whose relationships remain unresolved. Any system that claims to capture, preserve, or reconstruct "thoughts" is making a claim that the underlying science does not currently license — including any such claim made by Origo.

Everything below should be read against that finding.

Evidence classification

Every substantive claim in this document carries one of five labels. They describe the state of the evidence, not the interest of the idea.

LABEL	MEANING
ESTABLISHED	Replicated across labs and methods; broadly accepted within the relevant field. Disagreement is about detail, not existence.
STRONG EVIDENCE	Substantial converging support, but the account is still being refined and significant qualifications apply.
ACTIVE DEBATE	Multiple credible explanations remain live. Citing one as settled would misrepresent the field.
FRONTIER	Promising and actively researched, with limited, early, or narrow evidence.
SPECULATIVE	Conceptually interesting; not supported strongly enough to carry a scientific or product claim.

A label describes a claim, not a field. Memory science contains ESTABLISHED findings and SPECULATIVE ones simultaneously.

Layered map

Layer 1 — Phenomena, and how confidently they are characterized

PHENOMENON	BEST-SUPPORTED CHARACTERIZATION	CONFIDENCE
Perception	Constructive inference over sensory input, not passive transduction	ESTABLISHED
Attention	Multiple dissociable selection mechanisms, not one faculty	ESTABLISHED
Working memory	Limited-capacity maintenance and manipulation of information	ESTABLISHED
Long-term memory	Multiple systems (episodic, semantic, procedural) that dissociate	ESTABLISHED
Language	Structured combinatorial system, partially separable from general reasoning	STRONG EVIDENCE
Reasoning	Mix of fast associative and slow deliberative processing	STRONG EVIDENCE
Mental imagery	Quasi-perceptual representation with large individual variation	STRONG EVIDENCE
Inner speech	Common but variable internal verbal activity	STRONG EVIDENCE
Metacognition	Partially dissociable monitoring of one's own cognition	STRONG EVIDENCE
Spontaneous thought	Mind-wandering as a describable state with neural correlates	STRONG EVIDENCE
Concept formation	Category knowledge built from multiple representational formats	ACTIVE DEBATE
Consciousness	No agreed mechanistic account	ACTIVE DEBATE
"A thought" as a discrete unit	No agreed individuation criteria	ACTIVE DEBATE

Layer 2 — Candidate mechanisms

MECHANISM	WHAT IT IS PROPOSED TO EXPLAIN	CONFIDENCE
Neural population dynamics	Task-relevant computation as trajectories in population state space	STRONG EVIDENCE
Persistent and activity-silent maintenance	How information is held over seconds	ACTIVE DEBATE
Attentional selection and gain modulation	Why some information is prioritized	ESTABLISHED

MECHANISM	WHAT IT IS PROPOSED TO EXPLAIN	CONFIDENCE
Hippocampal pattern completion	How partial cues trigger full episodic retrieval	STRONG EVIDENCE (largely rodent)
Systems consolidation	Reorganization of memory across hippocampus and cortex over time	STRONG EVIDENCE
Synaptic plasticity / engram ensembles	Physical substrate of a stored memory	ESTABLISHED in rodents; FRONTIER in humans
Semantic activation across distributed cortex	How meaning is represented	STRONG EVIDENCE
Hub-and-spoke semantic organization	Integration of modality-specific features via anterior temporal hubs	STRONG EVIDENCE
Predictive coding	Perception as prediction-error minimization	ACTIVE DEBATE
Global broadcasting / ignition	Why some content becomes reportable	ACTIVE DEBATE
Recurrent local processing	Whether posterior recurrence suffices for experience	ACTIVE DEBATE
Reinforcement-learning value computation	How choices are evaluated	STRONG EVIDENCE
Reconsolidation	Whether retrieval renders memory labile and rewritable	ACTIVE DEBATE

Layer 3 — Fields and what each can settle

FIELD	STUDIES	CAN ESTABLISH	CANNOT ESTABLISH
Cognitive psychology	Behavioral signatures of processing	Functional dissociations, capacity limits	Neural implementation
Cognitive neuroscience	Brain–behavior relationships	Anatomical and temporal correlates	That a correlate is the thought
Systems / cellular neuroscience	Circuits, ensembles, plasticity	Causal mechanism in animal models	Direct translation to human semantics
Memory science	Encoding, retention, retrieval	Conditions that change what is remembered	The content of a specific human memory from outside
Psycholinguistics	Language processing	Structure and timing of comprehension and production	Whether language is required for thought
Computational cognitive science	Formal models of cognition	Sufficiency (a model can produce the behavior)	Necessity (that the brain does it that way)
Philosophy of mind	Concepts and their coherence	Whether a question is well-posed	Empirical facts
Clinical and lesion neuropsychology	Deficits after damage	Necessity of a region for a function	Sufficiency, or normal mechanism

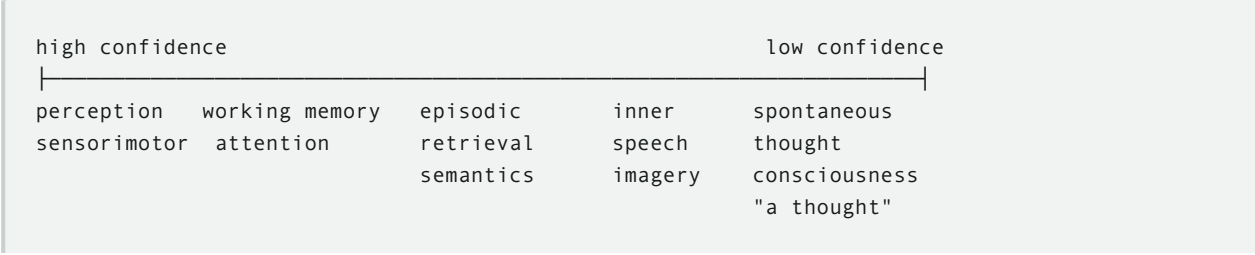
FIELD	STUDIES	CAN ESTABLISH	CANNOT ESTABLISH
Neuroengineering / BCI	Decoding and stimulation	What is recoverable from a signal under stated conditions	That the same is recoverable without cooperation

Layer 4 — Methods and their limits

METHOD	STRENGTH	STRUCTURAL LIMITATION
Behavioral experiment	Direct measure of performance	Underdetermines internal mechanism
fMRI	Whole-brain spatial coverage	Seconds-scale haemodynamic proxy; poor temporal resolution
EEG	Millisecond timing, low cost	Poor spatial localization; surface-biased
MEG	Millisecond timing, better localization than EEG	Expensive; still an inverse problem
Intracranial EEG / single-unit	High spatial and temporal precision	Only in clinical populations, at clinically determined sites
Lesion studies	Causal necessity claims	Rare, uncontrolled, subject to reorganization
Optogenetics / engram tagging	Causal manipulation of specific ensembles	Animal models only
Computational modeling	Precision, and forces assumptions explicit	Fit does not demonstrate mechanism
Psychophysics	Rigorous perception–judgment mapping	Narrow scope
Experience sampling	Access to spontaneous, real-world thought	Relies on introspective report
Longitudinal studies	Change over time	Confounded; slow; expensive

Layer 5 — Where confidence concentrates

Confidence is highest where the phenomenon is **behavioral, replicable, and close to sensation or motor output**. It falls sharply as the object of study becomes more internal, more abstract, more spontaneous, and more person-specific.



The practical consequence: the phenomena most relevant to a personal knowledge system — spontaneous idea formation, the felt sense of an unfinished thought, why something resurfaces when it does — sit at the **low-confidence end**.

Domains

1. Cognitive science

Accepted. Cognition involves internal states that mediate between stimulus and response, and those states carry information — the anti-behaviorist claim that founded the field. **ESTABLISHED.** Working memory is capacity-limited, and the limit is small — roughly three to four items for visual working memory, rather than the popular "seven plus or minus two," which described a different paradigm. **ESTABLISHED.** Attention is not a single faculty: spatial, feature-based, object-based, and executive attention dissociate behaviorally and neurally. **ESTABLISHED.** Expertise reorganizes representation rather than merely accelerating it — experts chunk differently, not just faster. **ESTABLISHED.**

Contested. The nature of mental representation is the field's oldest live dispute. Whether cognition operates over symbolic, structured representations or over distributed sub-symbolic ones — or requires both at different levels — remains unresolved after five decades. **ACTIVE DEBATE.** Cognitive architectures (ACT-R, Soar, and successors) demonstrate that integrated accounts of cognition can be built and can fit human data, but architecture fit does not demonstrate that human cognition is organized that way. **ACTIVE DEBATE.**

Relevant caution. Ego depletion, one of the most cited findings about cognitive control, largely failed to replicate in large preregistered multi-lab efforts. It remains widely quoted in popular writing about willpower and decision fatigue. Treat any single striking cognitive finding from before roughly 2015 as unverified until checked. **ESTABLISHED** that the replication problem is real; **ACTIVE DEBATE** on how much of the classic literature survives it.

2. Cognitive neuroscience

Accepted. Distinct large-scale networks are reliably associated with distinct cognitive modes — notably the frontoparietal control network with externally directed task performance and the default mode network with internally directed, self-referential, and mind-wandering states. **ESTABLISHED.** Medial temporal structures are necessary for forming new episodic memories. **ESTABLISHED.** Multivariate pattern analysis can decode which of a constrained set of stimulus categories a person is viewing. **ESTABLISHED.**

The critical caution. Complex thoughts do not map cleanly onto single brain regions. The mapping is many-to-many: one region participates in many functions, and one function recruits many regions. Reverse inference — reading "the amygdala was active, therefore the person felt fear" — is a known inferential error, not a conservative reading. The finding that a region is necessary (from

lesions) and the finding that it is active (from imaging) are different findings, and neither establishes that the region implements the function.

Contested. How working memory is maintained is genuinely open: persistent firing, activity-silent synaptic traces, and oscillatory mechanisms all have support, and they may coexist. **ACTIVE DEBATE.** Whether self-referential cognition constitutes a distinct kind of processing or a content-specific case of ordinary semantic and episodic machinery is unresolved. **ACTIVE DEBATE.**

3. Memory science

Accepted. Memory is not one system. Episodic (events), semantic (facts), procedural (skills), and working memory dissociate under damage and in normal function. **ESTABLISHED.** Encoding depends on depth of processing and attention at the time of the event. **ESTABLISHED.** Consolidation occurs over hours to years and depends substantially on sleep. **ESTABLISHED.** Retrieval is reconstructive, not reproductive: remembering assembles a plausible account from partial traces plus current knowledge, schema, and context. **ESTABLISHED.** The spacing effect and the testing effect are among the most robust findings in psychology — distributed practice and retrieval practice both substantially outperform massed restudy. **ESTABLISHED.** Retrieval cues matter enormously; much apparent forgetting is retrieval failure rather than trace loss. **ESTABLISHED.** Memories can be confidently held and factually wrong, and confidence is a poor guide to accuracy under many conditions. **ESTABLISHED.**

Contested. Reconsolidation — the claim that retrieval returns a memory to a labile state in which it can be modified or blocked — is well-demonstrated in rodents but has an inconsistent human literature, with notable replication failures and unclear boundary conditions governing when destabilization occurs at all. It is frequently overstated in popular and therapeutic writing. **ACTIVE DEBATE.** Engram research has established in rodents that specific neuronal ensembles can be tagged, and that reactivating them can drive recall and even induce false associations — a genuinely causal result. Whether and how this translates to human autobiographical memory is **FRONTIER**, limited by the fact that the enabling methods are invasive and animal-only.

Memory versus thought. They are distinguishable and interacting. Memory is the retention and reinstatement of prior states; thought is the current manipulation of representations. But retrieval is a form of thinking — episodic retrieval and imagining the future recruit substantially overlapping networks, which is why remembering and imagining are hard to separate cleanly. **STRONG EVIDENCE.** A stored record is not a memory: a record does not change on retrieval, and a memory does.

4. Inner speech

Accepted. Most people report internal verbal activity, and it correlates with verbal working memory performance. **STRONG EVIDENCE.** Inner speech engages speech-production regions, and the inner speech of people who experience auditory verbal hallucinations shows atypical monitoring signatures. **STRONG EVIDENCE.** Inner speech is not uniform: it varies along dimensions including condensation (compressed and telegraphic versus fully articulated) and dialogicality (monologue versus internal conversation). **STRONG EVIDENCE.**

Contested. Individual variation is much larger than usually assumed. Nedergaard and Lupyan (2024) proposed "anendophasia" for a reported absence of inner speech and found associated differences in verbal working memory and rhyme judgment. A 2025 commentary by Lind disputed the strong reading, arguing the evidence supports variation in frequency rather than genuine absence. **ACTIVE DEBATE.**

Limits of current knowledge. Nearly all inner-speech research depends on introspective self-report, which is exactly the measure whose reliability is under question in the metacognition literature. Descriptive Experience Sampling work suggests that people's global self-characterizations of their inner experience often diverge from moment-sampled reports. The field's central measurement instrument is therefore of uncertain validity, which is a structural limitation rather than a temporary one.

5. Mental imagery

Accepted. Visual imagery engages visual cortex, with content-specific patterns partially overlapping those of actual perception. **STRONG EVIDENCE.** Motor imagery engages motor planning regions and is used in rehabilitation and in BCI control paradigms. **STRONG EVIDENCE.** Imagery supports memory: mental imagery instructions reliably improve retention. **ESTABLISHED.**

Individual variation. Aphantasia — an absence or near-absence of voluntary visual imagery — is real and has objective corroboration beyond self-report, including differences in imagery-induced pupil responses and in physiological responses to imagined threat. Recent international prevalence estimates place aphantasia near 1%, hypophantasia near 3%, and hyperphantasia near 6%, with replication supporting those figures. **STRONG EVIDENCE.** People with aphantasia often retain intact semantic and spatial knowledge, which implies imagery is not required for the knowledge itself. **STRONG EVIDENCE.**

Implication worth holding onto. Two people can hold the same knowledge in radically different phenomenal formats. Any system that assumes a common inner format across users is assuming something false.

6. Metacognition

Accepted. Metacognitive sensitivity — the correspondence between confidence and accuracy — is measurable and dissociable from task performance itself: two people can perform identically while differing in how well their confidence tracks their correctness. **STRONG EVIDENCE.** Metacognitive ability is domain-specific to a substantial degree; good calibration in perception does not guarantee good calibration in memory. **STRONG EVIDENCE.** Prefrontal regions, particularly anterior prefrontal cortex, are implicated in confidence computation. **STRONG EVIDENCE.**

Contested. Introspective accuracy is the deep problem. Nisbett and Wilson's classic demonstration that people confabulate reasons for their own behavior has been refined but not overturned: people have limited access to the causes of their own judgments and readily generate confident but

incorrect explanations. How far this generalizes — whether introspection is unreliable about causes while reliable about contents — remains open. **ACTIVE DEBATE.**

Does metacognition bridge behavior and subjective thought? Partly, and it is the best available bridge. Confidence reports are behavioral, quantifiable, and trial-linked, which makes subjective states tractable in a way that free introspective report is not. But the bridge is narrow: it gives access to how certain someone is, not to what they are thinking. Treating metacognitive measurement as a route to thought content would overreach the method. **FRONTIER** as a general bridge; **STRONG EVIDENCE** in its narrow confidence-calibration form.

7. Predictive processing

Three claims are routinely conflated and should be separated.

Predictive coding as a specific mechanism. A hierarchical scheme in which higher levels predict lower-level activity and only the residual error propagates upward. There is real evidence for prediction-error signals: sensory responses are reliably attenuated for predicted stimuli, mismatch responses are robust, and reward prediction error in dopaminergic systems is one of the best-supported computational findings in neuroscience. **STRONG EVIDENCE** for prediction-error signaling; **ACTIVE DEBATE** on whether canonical predictive coding is the general cortical algorithm.

Predictive processing as a broad framework. The claim that perception, action, attention, and cognition are all fundamentally prediction. Explanatorily attractive and productive of experiments, but broad enough that disconfirmation is difficult. **ACTIVE DEBATE.**

The free energy principle and active inference. Friston's formulation is explicitly presented by its author as a mathematical principle rather than an empirical hypothesis — comparable to a principle of stationary action. Critics argue that this makes it unfalsifiable and explanatorily empty as neuroscience, since almost any observation can be reframed post hoc as free-energy minimization, and that it conflates thermodynamic and information-theoretic senses of free energy. **SPECULATIVE** as an empirical claim about brains; it may nonetheless be a useful modeling formalism, which is a different kind of value.

Predictive processing is not settled neuroscience. It is a strong programme containing some well-supported specific mechanisms.

8. Global workspace and conscious access

What it claims. Global Workspace Theory (Baars) and its neuronal formulation (Dehaene, Changeux) hold that content becomes conscious when it is selected and broadcast widely across a distributed network, particularly involving prefrontal-parietal regions, making it available to many consumer systems — report, memory, planning. The signature is "ignition": a late, nonlinear, widespread activation around 300 ms after stimulus for reportable content.

Evidence. Late widespread activity does reliably distinguish reported from unreported stimuli, and the P3b is a robust marker of conscious access. **STRONG EVIDENCE** for the empirical signature.

Criticism. The central objection is that the theory may be measuring reportability rather than experience: the late frontal activity may reflect the act of reporting, not the fact of being conscious. No-report paradigms — designed to isolate consciousness from the task of reporting it — have weakened some frontal findings. **ACTIVE DEBATE.**

9. Higher-order and other consciousness theories

Higher-Order Thought theories hold that a state is conscious when it is the object of a suitable higher-order representation. Elegant and connected to prefrontal metacognitive findings, but faces the problem of misrepresentation (what if the higher-order state is wrong?) and struggles with animal and infant consciousness. **ACTIVE DEBATE.**

Recurrent Processing Theory (Lamme) holds that local recurrent processing in sensory cortex suffices for phenomenal experience, independent of global access. Predicts consciousness without reportability, which is difficult to test by construction. **ACTIVE DEBATE.**

Integrated Information Theory (Tononi) identifies consciousness with integrated information (Φ) in a system's causal structure. It is unusual in proceeding from phenomenological axioms to physical postulates rather than the reverse, and it entails a form of panpsychism. In September 2023, 124 scholars signed an open letter arguing IIT should be labeled pseudoscience, primarily on grounds of panpsychist commitments and untestability of the theory as a whole. The letter itself became controversial: Anil Seth among others argued that strange or untestable consequences do not make a theory pseudoscientific if other components are testable, and surveys found only a minority of the field endorsing the label. **ACTIVE DEBATE**, on both the theory and its status.

The adversarial collaboration. The COGITATE consortium ran a preregistered adversarial test of GNWT against IIT, published in Nature in April 2025, with 256 participants across fMRI, MEG, and intracranial EEG, with predictions committed in advance by theory proponents. Results were **mixed and partially falsifying for both**. Some conscious content (stimulus category) was decodable from prefrontal cortex, but other consciously experienced features — orientation and identity — were not, and GNWT's predicted "ignition" burst was absent when conscious experience ended. Both are problems for GNWT. On the other side, IIT's predicted sustained synchronization between early and mid-level posterior visual areas was not observed, a direct challenge to the posterior "hot zone" claim. Neither theory was confirmed. This is the most methodologically serious test the field has run, and its outcome is that both leading theories require revision.

The honest summary: **there is no established scientific theory of consciousness.** There are competing frameworks with partial empirical support and known problems.

10. Concept formation and semantic knowledge

Accepted. Semantic memory is organized and is not reducible to episodic memory — semantic dementia and amnesia dissociate. **ESTABLISHED.** Typicality effects are robust: a robin is verified

as a bird faster than a penguin. **ESTABLISHED.** Category structure is graded, not classically definitional — most natural categories have no set of necessary and sufficient features. **ESTABLISHED.** Semantic knowledge is distributed across cortex, with a substantial body of work supporting a hub-and-spoke organization in which modality-specific features across cortex are integrated via anterior temporal hubs. **STRONG EVIDENCE.**

Contested. Prototype versus exemplar versus theory-based accounts of category representation have competed for four decades without resolution; current consensus is closest to "multiple representational formats coexist and which dominates is task-dependent." **ACTIVE DEBATE.** How abstract concepts (justice, irony, next quarter) are represented is much less understood than how concrete concepts are. **FRONTIER.**

Distributional representations. Word embeddings and language-model representations predict neural responses to language substantially better than earlier feature-based models — a real and replicated result. What it means is disputed: whether it shows the brain uses distributional statistics, or merely that distributional statistics are a good proxy for whatever the brain uses. Predicting a signal is not explaining it. **STRONG EVIDENCE** for the predictive result; **ACTIVE DEBATE** for the interpretation.

11. Reasoning and decision science

Accepted. Human judgment departs systematically from normative models in reproducible ways. **ESTABLISHED.** Reward prediction error signals in dopaminergic systems correspond closely to temporal-difference learning terms — among the most successful bridges between a computational model and a neural signal. **ESTABLISHED.** Value is compared in a common currency across categories, with orbitofrontal and ventromedial prefrontal involvement. **STRONG EVIDENCE.** Model-based and model-free reinforcement learning both contribute to human choice. **STRONG EVIDENCE.**

Substantially revised. Dual-process theory has changed more than its popular reception suggests. Evans and Stanovich, its principal proponents, retreated from "System 1 / System 2" language toward "Type 1 / Type 2" processing, because the systems framing wrongly implied discrete, modular, well-bounded subsystems. De Neys and colleagues went further with "dual process 2.0," arguing that intuitive processing frequently produces correct logical responses without deliberative intervention — which inverts the classic picture of fast-intuition-as-error-source. Popular business and design writing still uses the 2011-era version. **ACTIVE DEBATE** on architecture; **ESTABLISHED** that the simple two-systems picture is inadequate.

Several specific biases have weaker support than their citation counts imply; the general phenomenon of systematic bias is not in doubt, but individual effect sizes and moderators frequently are.

12. Embodied and situated cognition

The label covers claims of very different strength, and conflating them is the main error in this area.

Well supported. Sensorimotor systems are recruited during conceptual processing — reading action words engages motor regions in a somatotopically organized way. **STRONG EVIDENCE.** Cognition offloads onto the environment: people use gesture, notation, spatial arrangement, and external artifacts to reduce internal load, and disrupting these degrades performance. **ESTABLISHED.** Physical context affects memory retrieval. **STRONG EVIDENCE.**

Contested. The strong claim — that conceptual content is constituted by sensorimotor simulation, such that concepts cannot be represented without it — is not established. Patients with severe motor impairment retain action concepts, and abstract concepts resist the account. Whether motor activation is causal or a downstream correlate of comprehension remains disputed. **ACTIVE DEBATE.**

Speculative. Extended mind claims — that cognition literally includes external artifacts as constitutive parts rather than as tools — are primarily philosophical positions. They may be productive framings for thinking about personal knowledge infrastructure, but they are not empirical findings, and using them to justify a claim about what a product is would be a category error. **SPECULATIVE** as science.

13. Social cognition

Accepted. Humans routinely attribute mental states to others, and this ability follows a developmental trajectory with false-belief understanding typically consolidating around ages four to five. **ESTABLISHED.** A consistent network — temporoparietal junction, medial prefrontal cortex, precuneus, superior temporal sulcus — is engaged during mental-state attribution. **STRONG EVIDENCE.** Memory is socially shaped: conversational retelling alters what is subsequently remembered, including induced forgetting of unmentioned related material, and collaborative recall shows both costs and benefits relative to individual recall. **STRONG EVIDENCE.** Transactive memory — groups distributing responsibility for remembering across members and relying on each other as retrieval routes — is a robust finding. **STRONG EVIDENCE.**

Contested. Whether infants possess genuine theory of mind, based on implicit looking-time measures, is disputed following notable replication difficulties with implicit false-belief paradigms. **ACTIVE DEBATE.** The mirror-neuron account of social understanding, popular in the 2000s, is now considerably more contested than its public profile suggests. **ACTIVE DEBATE.**

Why it matters here. Thought develops substantially in dialogue and within relationships, not solely inside isolated skulls. Vygotskian accounts treating inner speech as internalized social speech remain influential and are broadly consistent with developmental evidence, though specific mechanisms are underspecified. **STRONG EVIDENCE** for social shaping of cognition; **ACTIVE DEBATE** for the specific internalization mechanism.

14. Language and thought

Accepted. Thought does not require language. Preverbal infants, nonhuman animals, and people with global aphasia demonstrate substantial reasoning without linguistic capacity — a well-supported line of evidence indicating that the language network is substantially dissociable from

general reasoning networks. **STRONG EVIDENCE.** The strong Whorfian claim that language determines thought is rejected. **ESTABLISHED.**

Well supported. Weaker linguistic-relativity effects are real: language influences perceptual discrimination at category boundaries (notably in color), spatial reference frames, grammatical-gender-linked associations, and event construal. These effects are genuine but typically modest and often task-dependent, appearing most reliably when language is available to be recruited during the task. **STRONG EVIDENCE.** Exact large-number reasoning depends on counting vocabulary — among the clearest demonstrations that a linguistic tool can enable a cognitive capacity. **STRONG EVIDENCE.**

Contested. Whether language is primarily a communication system that secondarily serves thought, or is itself a core medium of reasoning, remains disputed. Some correlational cross-linguistic findings, including the widely-cited future-tense-and-savings result, are confounded by cultural and economic factors and should be treated cautiously. **ACTIVE DEBATE.**

The claim "humans think in language" is false as a general statement. The defensible claim is that language is one powerful tool that thought can recruit, that people vary in how much they recruit it, and that its absence changes but does not eliminate thinking.

15. Computational cognitive science

What the models are. Bayesian models treat cognition as probabilistic inference under uncertainty. Reinforcement-learning models treat it as value learning from feedback. Symbolic models treat it as rule-governed manipulation of structured representations. Connectionist and neural-network models treat it as learned distributed transformation. Cognitive architectures attempt integration across the above.

What they establish. That a proposed mechanism is sufficient to produce a behavioral pattern, and that assumptions are explicit enough to be tested. Reinforcement learning in particular achieved a genuine mechanistic bridge to dopaminergic signaling. **ESTABLISHED** as a methodology.

What they cannot establish. That the brain implements the model. Model fit is weak evidence for mechanism, because many models fit the same data, and flexible models fit almost anything. Bayesian accounts are especially prone to this: with freedom to choose priors and likelihoods, near-any behavior can be rendered "optimal" under some specification, and the framework has been criticized on exactly these grounds. **ACTIVE DEBATE** on how much Bayesian fit tells us.

The large-language-model question. LLMs produce fluent language and predict neural language responses well. Whether this indicates shared computational principles with human language processing, or is a case of very different systems converging on similar statistical structure, is genuinely open and currently among the field's most consequential unresolved questions. **FRONTIER.** Treating LLM behavior as evidence about human cognition is not currently warranted.

16. Neurotechnology and thought decoding

This section requires the most precision, because it is the area where popular reporting diverges most from the evidence.

Invasive methods

Intracranial recording in clinical populations has produced the field's strongest results. Speech neuroprostheses now decode attempted speech in people with paralysis at conversational rates with large vocabularies and substantially reduced word error rates — a genuine clinical achievement, and one that restores communication to people who have lost it. **ESTABLISHED** as a capability under its stated conditions.

The most relevant recent result: Kunz, Krasa, Willett and colleagues at Stanford published in Cell in August 2025 that inner speech — imagined, not attempted — is robustly represented in motor cortex and can be decoded in real time, reaching up to 74% accuracy on sentences drawn from a 125,000-word vocabulary in four participants with ALS or brainstem stroke. This is the strongest evidence to date that imagined speech is decodable. **FRONTIER**, with important qualifications: four participants, all with implanted electrodes, all deliberately generating inner speech on cue, in a laboratory, with per-participant trained decoders, at accuracy well below transcription quality.

Notably, the same researchers treated mental privacy as a design problem rather than an afterthought: they identified a neural signal distinguishing attempted from inner speech, and implemented a keyword-based unlocking mechanism so that inner speech is only decoded when the user intends it. This is a meaningful precedent — the field's own leading practitioners regard uninvited decoding as something to be engineered against.

Non-invasive methods

Tang, Huth and colleagues (Nature Neuroscience, 2023) demonstrated continuous semantic decoding from fMRI: a decoder that recovers the gist of perceived speech, imagined speech, and even silent film, rather than selecting from a small closed set. This was a real advance. Its constraints are as important as its result:

- It requires roughly 16 hours of individually collected training data per person.
- Decoders are subject-specific and transfer poorly across people.
- Output is paraphrase-level gist, not verbatim recovery.
- fMRI requires a stationary participant in a large, expensive scanner.
- **Decoding requires subject cooperation.** The authors ran an explicit resistance experiment and found participants could defeat the decoder with simple mental strategies. They withheld that data from public release specifically to avoid enabling circumvention of resistance.

Visual reconstruction from fMRI using diffusion models has improved markedly, with training-data requirements falling from roughly forty hours to about one hour per subject in shared-subject models. Reconstructions are often visually compelling while being unfaithful to the actual stimulus in structure and semantic detail, and constructing a model for a new subject generally still requires retraining. **FRONTIER.**

EEG decoding is constrained by poor spatial resolution and works best for well-trained, coarse categorical distinctions. MEG decoding is better and improving, with self-supervised approaches beginning to show cross-subject generalization, but remains far from free-form thought recovery. **FRONTIER.**

Direct answer to the question

> **Can science currently read human thoughts non-invasively?**

No — not in the sense the phrase implies. More precisely:

1. **Not without cooperation.** Every non-invasive result requires the person to sit still for many hours of individualized decoder training and then to cooperate during decoding. Resistance defeats current decoders.
2. **Not without per-person training.** Decoders do not transfer well across individuals. There is no general-purpose reader.
3. **Not verbatim.** Non-invasive language decoding recovers gist, not exact inner wording.
4. **Not free-form.** Results come from constrained paradigms — listening to known stories, viewing images from known distributions, generating cued sentences. Spontaneous, unprompted, unconstrained thought has not been decoded.
5. **Not portable.** The best non-invasive results require an fMRI scanner.
6. **Yes to something real and narrower.** Semantic gist from cooperating, extensively trained individuals in a scanner is genuinely recoverable. That is a meaningful scientific result and a legitimate reason for neuro-rights attention, and it should not be dismissed.

The trajectory matters even though the current state is limited. Invasive inner speech decoding is advancing quickly, foundation models for neural data are reducing per-subject data requirements, and non-invasive sensing continues to improve. The correct posture is neither dismissal nor alarm: this is a real research direction whose current capabilities are narrow and whose ethical questions are already being taken seriously by its own practitioners.

Media claims describing "mind-reading AI" generally omit the cooperation requirement, the per-subject training, the constrained stimulus set, and the gist-versus-verbatim distinction. Assume any such headline is overstated until the paper's methods section confirms otherwise.

Major open questions

These are unresolved. Several may be badly posed rather than merely unanswered, which is itself a finding.

1. **What constitutes a single thought?** No discipline has non-arbitrary individuation criteria. There is no agreed answer to where one thought ends and the next begins, and it is unclear whether the question is well-formed.

2. **How are concepts physically represented?** Distributed patterns predict neural responses, but what the representation is — and whether "the concept" is a stable thing or a context-dependent reconstruction — is open.
3. **How does spontaneous thought begin?** Mind-wandering has neural correlates, but what initiates a specific unprompted thought at a specific moment is essentially unknown.
4. **How is content selected for conscious awareness?** The central question the competing consciousness theories exist to answer, and the COGITATE result suggests none currently answers it adequately.
5. **How reliable is introspection?** People confabulate causes confidently. How far this extends to reports of thought content — the basis of nearly all self-report methodology — is unresolved and methodologically severe.
6. **How does autobiographical memory shape current reasoning?** Retrieval and imagination overlap neurally, but the causal path from remembered episodes to present judgment is not well characterized.
7. **What is the relationship between memory retrieval and imagination?** Substantial overlap is established; whether they are one capacity applied to different targets is not.
8. **Can inner speech be reliably decoded?** Early invasive evidence says partially, in cued laboratory conditions. Spontaneous inner speech in ordinary life is untested.
9. **How distributed is semantic representation, and how idiosyncratic?** Cross-subject alignment works better than chance and worse than would be needed for a general decoder. Where the boundary sits between shared and personal semantic structure is open.
10. **How does emotion alter thought formation?** Emotion clearly modulates encoding, retrieval, and judgment; the mechanisms by which affect shapes which thought arises are poorly specified.
11. **How does social context shape cognition?** Established that it does; unresolved how much of individual reasoning is constitutively social.
12. **What distinguishes conscious from unconscious processing?** Competing accounts, no resolution, and disagreement about what evidence would settle it.
13. **How are long-term goals represented across time?** How an intention persists across months, remains available for reactivation, and shapes intervening cognition is not well understood.
14. **Can neural decoding generalize across individuals?** Currently poorly. Whether foundation models overcome this or hit a genuine idiosyncrasy ceiling is an open empirical question.
15. **Is there a common format across modalities of thought?** Aphantasia suggests the same knowledge can exist in radically different phenomenal formats, raising the question of what is preserved across those differences.
16. **What makes a thought feel unfinished?** The experience of an incomplete idea that later resolves is nearly universal and barely studied.
17. **How does a person recognize that a past thought is relevant now?** The cue-to-retrieval relationship is well studied in the laboratory; the real-world case of an idea resurfacing appropriately is not.

18. **Does language change conceptual structure, or only access to it?** Weak relativity effects are established; whether they reflect altered concepts or altered strategy remains disputed.

Frontier research

Semantic decoding from brain activity

Aim. Recover meaning, rather than a category label, from neural signal. **Evidence.** Real for perceived and imagined speech and silent film under fMRI with heavy per-subject training. **Limits.** Cooperation-dependent, subject-specific, gist-level, scanner-bound. **Near term.** Reduced training requirements; better cross-subject transfer; more naturalistic stimuli. **Breakthrough would be.** Reliable decoding of spontaneous thought from a new person with no individualized training. Nothing currently indicates this is close. **FRONTIER.**

Imagined and inner speech decoding

Aim. Let people communicate by thinking rather than by attempting speech. **Evidence.** Stanford's 2025 Cell result: real-time inner-speech decoding, 74% accuracy, large vocabulary, four implanted participants. **Limits.** Invasive; tiny n; cued rather than spontaneous; per-participant decoders. **Near term.** Higher accuracy, more participants, better separation of intended from unintended inner speech. **Breakthrough would be.** Sustained accurate decoding of self-initiated inner speech with a reliable user-controlled gate. **FRONTIER**, moving fast.

Foundation models for neural signals

Aim. Pretrain across subjects, tasks, and modalities so that new subjects need little data. **Evidence.** Genuine progress — self-supervised MEG representations that scale with data and transfer across subjects and tasks; encoding accuracy improving log-linearly with no clear ceiling reached; models predicting responses to stimulus types absent from training. **Limits.** Heterogeneous recording setups, small aggregate data by machine-learning standards, unclear whether individual idiosyncrasy imposes a hard floor. **Breakthrough would be.** Demonstrated zero-shot decoding in an unseen subject. **FRONTIER.**

Memory reconstruction and engram translation

Aim. Understand and eventually influence specific memory traces. **Evidence.** Strong and causal in rodents; the enabling methods are invasive and species-limited. **Limits.** No human equivalent exists or is near. **Breakthrough would be.** A non-invasive human method for identifying a specific memory trace. Not currently foreseeable. **FRONTIER** in animals, **SPECULATIVE** in humans.

Neural representation alignment

Aim. Determine whether artificial and biological systems converge on shared representational geometry. **Evidence.** Language models predict neural language responses well above prior baselines — replicated. **Limits.** Prediction is not explanation; shared statistical structure in the input may suffice to explain convergence. **Breakthrough would be.** A causal intervention showing that a representational property in a model is the one the brain uses. **FRONTIER**, and conceptually contested.

Longitudinal personal cognition research

Aim. Study one person's cognition densely across months or years, rather than many people once. Precision-imaging work has shown that individual functional organization is stable, reliable, and meaningfully different from group averages. **Limits.** Expensive; tiny n; generalization unclear. **Breakthrough would be.** Demonstrating that individual-level cognitive trajectories predict something that group-level models cannot. This is plausibly the frontier most relevant to personal knowledge infrastructure, and it requires no neural access at all. **FRONTIER.**

Computational psychiatry

Aim. Recast psychiatric conditions as identifiable computational differences — altered priors, learning rates, precision weighting. **Evidence.** Some replicable group-level parameter differences. **Limits.** Effect sizes generally too small for individual-level clinical use; heavily dependent on contested predictive-processing framing. **Breakthrough would be.** A computational marker with genuine individual predictive validity. **FRONTIER.**

AI-assisted neuroscience

Aim. Use machine learning to find structure in neural data beyond hand-designed analyses. **Evidence.** Already productive as a method. **Limits.** Risk of increasingly accurate prediction with no gain in understanding — the field's central methodological worry. **FRONTIER** as a method; not a claim about cognition.

Possible Relevance to Origo

This section maps research areas to questions Origo may eventually face. It proposes no features, endorses no capability, and creates no product or architectural authority. Nothing here implies neural access, mind reading, or any form of thought capture. Every item is a question, and every item is labeled.

Near-term relevance

Reconstructive retrieval argues for preserving original expression. That human memory is reconstructive rather than reproductive is **ESTABLISHED**. People's recollection of what they thought drifts, and drifts confidently. This is the strongest scientific support in this document for something Origo already holds: an unaltered record of what a person actually expressed has value precisely because the person's own memory of it will not stay fixed. Question raised: how should Chronicle preserve original expression such that the drift between record and recollection is visible rather than silently resolved in favor of either? **NEAR-TERM RELEVANCE.**

Source-versus-interpretation separation has independent scientific support. Metacognitive research **STRONG EVIDENCE** shows people readily generate confident but incorrect accounts of their own reasoning. If human introspective report is already an interpretation layered over an underlying process, then blending a machine interpretation into the record compounds an existing epistemic problem rather than merely adding a new one. Question raised: does Origo's existing source/derived separation need to extend to distinguishing a person's later interpretation of their own earlier thought from the earlier expression itself? **NEAR-TERM RELEVANCE.**

Spacing, testing, and cue-dependence bear on resurfacing. The spacing effect, the testing effect, and cue-dependent retrieval are **ESTABLISHED**. They are among the most robust findings in psychology and require no neuroscience. Questions raised: if material is resurfaced, does timing matter, and on what evidence? Does resurfacing something as a prompt differ in effect from resurfacing it as a statement? Does re-encountering a preserved record change the memory of the original event, and is that acceptable? The last question is a real risk, not only an opportunity: a system that resurfaces records may alter the recollection it was meant to protect. **NEAR-TERM RELEVANCE.**

Confidence calibration is measurable, and mis-calibration is harmful. Metacognitive sensitivity is **STRONG EVIDENCE**. If Origo ever presents an inferred relationship, the honest representation of its uncertainty is not a presentation detail; it is the difference between supporting and degrading a person's own calibration. Question raised: how should uncertainty in inferred relationships be represented so that it is actually used rather than decoratively displayed? **NEAR-TERM RELEVANCE.**

People do not share an inner format. Aphantasia prevalence near 1% and hyperphantasia near 6% (**STRONG EVIDENCE**), plus contested but real variation in inner speech (**ACTIVE DEBATE**), mean any assumption of a common inner representational format across users is false. Question raised: what does Origo assume about how users internally represent their own material, and should it assume anything? **NEAR-TERM RELEVANCE.**

Long-term research

Metacognition as a design frame for reflective interfaces. Metacognition is the best-evidenced bridge between observable behavior and subjective thought, and it is narrow: it gives confidence, not content. Question raised: could a reflective surface support a person's own metacognition — helping them notice what they were uncertain about — without the system claiming to know what they thought? This is an open research question, not a feature. **LONG-TERM RESEARCH.**

Concept formation and knowledge linking. Graded category structure and distributed semantic representation are **ESTABLISHED** and **STRONG EVIDENCE** respectively. But the finding that natural categories lack necessary-and-sufficient definitions is a caution as much as an opportunity: a person's own conceptual organization is graded, context-dependent, and shifting. Question raised: what would it mean for a linking mechanism to respect that a person's categories are unstable, rather than imposing a fixed taxonomy? **LONG-TERM RESEARCH.**

Social and dialogic shaping of thought. Transactive memory and conversational shaping of recall are **STRONG EVIDENCE**. Thought develops between people, not only within them. Question raised: what does Personal Knowledge Infrastructure that takes relationships seriously look like, without compromising the privacy commitments that make it trustworthy? These pull against each other, and the tension is real. **LONG-TERM RESEARCH.**

Emotion, context, and time as retrieval structure. Context-dependent retrieval is **STRONG EVIDENCE**; the mechanisms by which emotion shapes which thought arises are **FRONTIER**. Question raised: how should Personal Knowledge Infrastructure account for emotional and temporal context without inferring emotional states it cannot verify? **LONG-TERM RESEARCH.**

Longitudinal individual cognition. Precision-imaging findings that individual functional organization is stable and distinct from group averages are **FRONTIER**, but the underlying idea — that studying one person densely beats studying many people once — has an obvious affinity with a durable personal record. Question raised: what could be learned from an individual's own longitudinal record of expressed thinking, with explicit consent, that population research cannot show? **LONG-TERM RESEARCH.**

Speculative

Anything involving neural signal. Origo has no relationship to neural decoding, and this document should not be read as establishing one. Decoding requires implanted electrodes or a research scanner, per-person training, and active cooperation. **SPECULATIVE**, and listed here only so that the boundary is explicit rather than assumed.

Extended-mind framings of Personal Knowledge Infrastructure. Philosophically interesting; **SPECULATIVE** as science. It may be a useful way to think. It is not evidence, and it cannot support a claim about what Origo is.

Reconstructing what a person thought from what they wrote. The record is the expression, not the thought. Given that no science individuates thoughts (**ACTIVE DEBATE** on whether the question is even well-posed), inference from text to underlying thought is interpretation and must remain labeled as such. **SPECULATIVE.**

What this research does not support

Stated explicitly, because absence of support is easy to lose:

- No finding supports inferring a person's beliefs, intentions, or emotional states from their records with accuracy sufficient to treat the inference as fact.

- No finding supports treating an AI-generated summary as equivalent to what a person thought.
- No finding supports the claim that a stored record preserves a memory. Records persist unchanged; memories do not.
- No finding supports any capability requiring neural access.
- No finding establishes what a thought is, which means no system can honestly claim to capture one.

Explicit Origo boundary

These distinctions are load-bearing. Each is grounded in the findings above, not merely asserted.

```
thought           ≠  Chronicle entry
                     A Chronicle entry is an expression a person chose to
                     record. Science cannot individuate thoughts at all.

memory            ≠  stored record
                     Retrieval is reconstructive and changes the trace.
                     A record is static. These are different kinds of thing.

neural activity   ≠  meaning
                     Correlated patterns decode constrained categories
                     under cooperation. That is not access to meaning.

correlation       ≠  explanation
                     Distributional models predict neural language responses
                     well. Prediction is not mechanism.

AI interpretation ≠  human thought
                     Inference over expression is a derived artifact and
                     must remain labeled as one.

decoded signal    ≠  unrestricted mind reading
                     Every result requires cooperation, per-subject training,
                     and constrained conditions. Resistance defeats decoders.

scientific model  ≠  human person
                     Every model here is a deliberate simplification.
                     None of them is a person, and none is Origo's user.
```

Review status and limitations

This document is a working research survey, not a decision. It creates no product, architectural, or foundational authority, and nothing in it should be cited as a basis for a capability claim.

Known limitations of the survey itself:

- It is a synthesis, not a systematic review. Domain coverage is uneven and reflects relevance to Personal Knowledge Infrastructure rather than proportional coverage of each field.

- Evidence classifications are judgments, not the output of a formal assessment protocol. They should be treated as arguable.
- Fast-moving areas — particularly neural decoding and neuro-foundation models — will date quickly. Re-verify anything in section 16 before relying on it.
- Psychology and neuroscience both have documented replication problems. Findings labeled ESTABLISHED here are those with multi-lab or multi-method support, but the label is a claim about current evidence, not a guarantee.

Sources

Primary sources consulted for time-sensitive claims:

- Tang et al., Semantic reconstruction of continuous language from non-invasive brain recordings, Nature Neuroscience (2023)
- Kunz et al., Inner speech in motor cortex and implications for speech neuroprostheses, Cell (2025)
- NIH Research Matters, Decoding inner speech from brain signals
- COGITATE Consortium, Adversarial testing of global neuronal workspace and integrated information theories of consciousness, Nature (2025)
- Nature news, Consciousness theory slammed as 'pseudoscience' (2023)
- Seth, The Worth of Wild Ideas, Nautilus
- Nedergaard & Lupyan, Not Everybody Has an Inner Voice, Psychological Science (2024)
- Lind, Are There Really People With No Inner Voice? Psychological Science (2025)
- An international estimate of the prevalence of differing visual imagery abilities, Frontiers in Psychology (2024)
- Stemerding et al., Demarcating the boundary conditions of memory reconsolidation: an unsuccessful replication, Scientific Reports (2022)
- MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data
- Foundation model of neural activity predicts response to new stimulus types, Nature (2025)
- Brain Foundation Models: A Survey on Advancements in Neural Signal Processing and Brain Discovery
- Toward dual-process theory 3.0, Behavioral and Brain Sciences
- Gershman & colleagues, What does the free energy principle tell us about the brain?
- Josselyn & Tonegawa, Memory engrams: Recalling the past and imagining the future, Science

---

Deep Well Labs, Inc. advances the science and infrastructure of human thought. Origo is a product of Deep Well Labs, Inc. This public edition is derived from the company's internal source document and may be revised. deepwelllabs.com · phil@deepwelllabs.com