Revista Română de Filosofie Analitică,
Romanian Journal of Analytic Philosophy
Vol. XVII, Iulie-Decembrie 2023, Nr. 2, p. 62-85
ISSN (ediția electronică): 1843-9969,
ISSN (ediția tipărită): 1844-2218
DOI: 10.62229/rrfaxvii-2/3

HOW DO WE GROUND 'GROUNDING'? WHY THE VECTOR GROUNDING PROBLEM REMAINS UNSOLVED

FLORIN COJOCARIU[1]

Abstract: On the view I develop, grounding is a phenomenological achievement of organisms — structured, concern-laden encounter with the world, observable in basic minds that connect to their environment *without* representational content. Language coordinates with this connection; it does not create nor implement it. I develop these views within the Pattern Recognition Unity (PRU) framework, distinguishing experiential Pattern-Constellations from Linguistic Pattern-Constellations and exploring how they coordinate. Contra Mollo and Milliere's (2025) view, I further argue that teleosemantics — a theory that earned its credibility on pre-verbal organisms selected against grounded functions like feeding and fleeing — loses its explanatory force when applied to systems selected against linguistic adequacy. The selection history of LLMs *borrows* its world-connection from human raters rather than *constituting* its own; its causal relation to *worldly features* is inherited ancestry, not the proximal coupling teleosemantics describes in the natural world. It follows that no evaluation conducted within the linguistic register can distinguish genuine grounding from sufficiently powerful linguistic mastery.

Keywords: *Grounding; LLMs; Consciousness; Explanatory Gap; Teleosemantics; Basic Minds.*

[1] Florin Cojocariu is a student in the 'Analytic Philosophy 'master programme within the Faculty of Philosophy at the University of Bucharest.

1. The Grounding Problem

1.1.The Standard Question

How does the *word*" cat" connect to actual cats? This question — the grounding problem in miniature — has generated three decades of debate in philosophy of mind, cognitive science, and now philosophy of AI. From Harnad's (1990) Symbol Grounding Problem and Searle's Chinese Room (Searle, 1980) to the current dispute over LLMs, the question tracks the same structure: on one side, a representation (word, symbol, internal state); on the other, the world (cats, fires, chairs); between them, a gap that the theory must bridge.

The problem persists because representations *qua* representations relate only to other representations. A word connects to other words. A definition connects to other definitions. An internal state described as "carrying information about cats" connects to *the description of cats*, not to *cats*. This is the representational merry-go-round: representations all the way down, that never reach the actual world they're representing.

The standard framing, then, treats grounding as *a property that representations need to acquire*. The word "cat" lacks something — a connection, a hook, a tether to the world — and the task is to identify what that something is and under what conditions it can be obtained.

1.2. The Best Available Disambiguation

(Mollo and Milliere, 2025) bring about progress by showing that "grounding" is tasked with too many jobs at once.[2] They distinguish five notions:

(i) Referential grounding: a representation connects to worldly entities or properties.

[2] For an anticipation of questions concerning how these different notions of grounding overlap, albeit in a different terminology, that of reference v. description for singular terms, cf. (Dumitru, 2004).

(ii) Sensorimotor grounding: a representation is connected to sensory or motor representations *inside the mind*.

(iii) Relational grounding: a representation is linked to other linguistic representations through inferential or distributional relations *inside the mind*.

(iv) Communicative grounding: representations are calibrated through linguistic practices between speakers to establish common ground.

(v) Epistemic grounding: a representation is linked to a knowledge base — a corpus of specific knowledge, formally or informally built (text or culture).

Of the five, only referential grounding direct hooks onto worldly entities. Mollo and Millière conclude, correctly, that this is the only notion relevant to the Vector Grounding Problem for LLMs. They argue that the other four either trade between representations alone, or presuppose the semantic content they would establish, or defer the question to a further structure that itself requires grounding.

The four non-referential notions are not failed *kinds* of grounding. They are concepts *about* grounding: linguistic descriptions, formulated in publicly articulable prose, of how representations might relate to *something*. Each names a relation between representations. Each can be stated, refined, debated. Each is a description rather than a direct hook to the thing described. Concepts about grounding are not grounding.

Referential grounding is supposed to be grounding itself, and that conditions: causal-informational relations, selection history, the function of tracking worldly features. Simply coming up with another concept about grounding — another publicly articulable description, connected inferentially to other descriptions in the teleosemantic literature—would only generate a fifth entry in Mollo and Millière's list.

The problem is structural: language can only produce concepts about grounding. It cannot produce grounding itself, because grounding, if real, is not a linguistic relation. Every attempt to specify what referential grounding is converts it into a description of what it is. The representational framework reinforces this limitation: representations are

items that bear content, content is evaluated for accuracy, accuracy is assessed in language. Everything the framework touches becomes a candidate for linguistic evaluation. It cannot describe the organism's pre-linguistic encounter with the world without converting it into a representation.

The debate cycles because it treats grounding as a property language must deliver, when grounding is something, organisms achieve and language coordinates with.

1.3. The Hidden Assumption

The standard framing contains a hidden assumption that makes the problem structurally unsolvable from within. The assumption is this: grounding is a property of representations — something language needs to acquire. The question "how does the word 'cat connect to actual cats?" places grounding in the linguistic domain. Its subject is the word. The word is what lacks grounding. The world sits inert, waiting to be hooked onto. The entire problematic runs from language outward: *symbol → ??? → world*.

Consider a cat tracking a mouse. The cat's mind connects to the mouse with precision: position, speed, trajectory are tracked; the body organizes around the encounter; salience and urgency radiate from the prey. No language is involved. No representations with truth conditions are in play. Yet the connection between mind and world is as tight as any the grounding debate discusses.

What should we call this? It is not grounding in the standard sense — there is no linguistic content to ground[3]. But it is a mind-world

[3] This aligns with the Radical Enactive Cognition (REC) programme of Hutto and Myin (2013), which argues that basic minds — minds without language — are world-directed but contentless. On their account, genuine semantic content with conditions of satisfaction emerges only when organisms participate in sociocultural linguistic practices. What REC does not provide is a positive structural account of what basic world-directedness consists

connection: a structured, stable, concern-laden engagement with a worldly entity, formed through the organism's history of encounter. Call it *basic grounding*: the pre-linguistic, pre-content connection between an organism and its world that the standard grounding debate presupposes.

Basic grounding is what makes linguistic grounding possible. The child who learns the word "cat" does not use the word to reach cats. She already reaches cats — through basic grounding built from world encounter-events. The word arrives and coordinates with a connection that is already in place. The child's accumulated cat-encounters — serial, *locus*-bound, laden with affect and concern — have formed a stable experiential pattern: an attractor basin integrating visual, tactile, auditory, motor, and affective dimensions of cat-encounter. This pattern is not a representation of cats. It is the sedimented structure of the organism's encounters with cats from its particular *locus* in the world. It exists before language. It exists in animals who will never have language. It is built through encounter, not through representation.

Grounding, properly understood, does not run from language to world. It runs from world to organism — basic grounding, the formation of experiential patterns through encounter — and then language coordinates with the result. The standard question inverts this order. By asking "how does the word connect to the world?" it implies that the problem begins with language and must be solved within language. But the organism is already on the world's side. The bridge language needs is not to the world but to the organism's experiential engagement with the world — an engagement that is subjective, pre-linguistic, and unavailable in the collective register of language.

---

in — what its units are and how it is organized. The experiential Pattern-Constellations ({X}) developed in Section 2 are offered as this structural unit: contentless but stable, phenomenologically real, empirically investigable. Content, on this account, emerges at $\mathcal{R}(\{X\},\langle x\rangle)$— in the coordination between experiential and linguistic patterns — which locates the emergence of content exactly where REC predicts: in the organism's entry into sociocultural-linguistic practice. PRU thus completes rather than competes with REC, providing the structural architecture for a claim REC makes but does not fully specify.

1.4. Grounding Redescribed

If grounding is not a property of representations but a phenomenological achievement of organisms, what does the grounding problem actually ask? Not: "How do words connect to the world?" But: "How does language coordinate with the experiential patterns organisms have already formed through worldly encounter?"

The first question is unanswerable within language because it locates grounding in language. The second question is tractable because it separates things that the first question conflates:

(a) The experiential pole: the organism's stable pattern of encounter, subjective, individual, formed pre-linguistically through serial, *locus*-bound, concern-laden events. This is what basic grounding produces. It is what the cat has with respect to the mouse and what the child has with respect to cats before she learns the word.

(b) The linguistic pole: the public, collective pattern of word-use, intersubjective, sustained through communal practice. This is what language *is*.

(c) Multifarious coordination: the learned couplings between (a) and (b) — the processes by which an individual's experiential patterns become coordinated with a community's word-use patterns. This is what reference *is*.

The grounding problem, in its standard form, collapses (a), (b), and (c) into a single question about representations. The result is that (a) — the subjective, phenomenological pole — becomes invisible. It is not denied; it is simply erased by a framing that has no place for it.

What follows? Evaluating whether LLMs are "grounded" requires a framework that can hold (a) and (b) apart — that can mark the experiential pole without converting it into a linguistic description, and mark the linguistic pole without letting it silently stand in for the whole. A meta-language is then needed: a notation equipped with devices that indicate the experiential pole without redescribing it, mark the linguistic pole as linguistic, and represent the coordination between them as a structured relation between different kinds of thing.

Mollo and Millière's argument is that LLMs can have internal states genuinely *about* the world — not just interpretable as such — provided those states reliably track worldly features and were shaped by a learning process (especially RLHF) that selected them for doing so. On their view, neither embodiment nor multimodality is required; the right kind of training history is enough. This is the strongest available case that LLMs achieve referential grounding and it operates entirely within (b), the linguistic pole. Their teleosemantic conditions describe (b)-type properties: distributional structure, functional roles, causal-chain descriptions. LLMs, which are (b)-type systems par excellence, naturally satisfy (b)-type conditions. The evaluation succeeds because evaluator and evaluated share the same register. The experiential pole (a) is not detected as missing because the framework has no way to detect it at all. This is not Mollo and Millière's fault. It is a structural consequence of asking a question about subjective experience using tools that can only describe collective, publicly articulable properties.

2. The PRU Notation as Meta-Language

The preceding section identified three things a meta-language must mark: (1) the experiential pole — subjective, individual, non-transmissible through text; (2) the linguistic pole — collective, public, transmissible through text; (3) coordination relations between them. I develop three corresponding notational devices within the Pattern Recognition Unity (PRU) framework[4].

{X} — Experiential Pattern-Constellation. The curly braces function as *notational quarantine*: they indicate the experiential pole without redescribing it in linguistic terms. The moment you write "the

[4] Cojocariu, Florin, Pattern-Recognition Unity (PRU): A Framework Specification A Meta-Language for Grounding, Reference, and the Experience–Language Interface (February 10, 2026). Preprint available at SSRN: https://ssrn.com/abstract=6285878 or http://dx.doi.org/10.2139/ssrn.6285878

experiential constellation of fire," you have produced a piece of language — a concept about {X}. {FIRE} resists this by pointing at the experiential pole the way an index finger points at a dog. It does not describe the dog. It does not deliver the dog. It marks what is experienced when encountering a dog, and the description stops here.

{X} is what the cat has with respect to the mouse — the stable, multi-modal, affect-laden attractor basin formed through the organism's history of encounter-events. It is what the previous section called basic grounding, now given a notational handle. {X} is always individual, always formed through encounter-events, always *locus*-bound. {FIRE} is *my* fire-constellation or *yours* — there is no view-from-nowhere {FIRE}.

⟨x⟩ — Linguistic Pattern-Constellation. The angle brackets mark the linguistic pole *as linguistic*. ⟨dog⟩ denotes the public pattern of how "dog" functions in language — its distributional profile, inferential relations, sentential regularities — and nothing more. The framing is broadly Wittgensteinian: ⟨dog⟩ is the meaning of "dog" insofar as meaning lives in the public practice of use.

⟨x⟩ is always collective, always public, always transmissible through text. ⟨dog⟩ is not mine or yours — it is the communal pattern of use. It is what LLMs learn.

$\mathcal{R}$({X}, ⟨x⟩) — The coordination relation. $\mathcal{R}$ marks that referential grounding is a *structured relation* between two different kinds of pattern. It is not a single hook from word to world. It is the learned coupling between an individual's experiential attractor basin and a collective pattern of word-use. Reference works because $\mathcal{R}$ holds — because my {DOG} and your {DOG}, despite being different, are each coordinated with the shared ⟨dog⟩.

$\mathcal{R}$ marks the coordination as a coordination — as a relation between two things the notation has formally separated. Without the separation, "reference" looks like a single, undifferentiated relation between word and world. With the separation, reference reveals itself as a bridge between something the notation can describe (⟨x⟩) and something it can only indicate ({X}).

Throughout this paper I'll adhere to the following simplifying convention: if context asks for it, by {X} I mean the totality of the pattern constellation an organism developed at a given time, something more precisely written as $\bigcup_i\{X_i\}$. By ⟨x⟩ we mean the totality of language pattern constellations a human possesses at the same moment, something more precisely written $\bigcup_i \langle x_i \rangle$.

The notation does not deliver {X} to the reader. No notation could. What it does is create *representational hygiene* — a discipline that prevents a specific conflation from occurring unnoticed: the conflation of the collective linguistic pole with the full structure of grounding.

At this stage, PRU is less than a formalism. A formalism derives results, generates theorems, performs calculations. PRU does not do that here. It is a meta-language: a language about the limitations of language.

But the notation is also more than a metaphor. The {X}/⟨x⟩ distinction has empirical traction and developmental predictions, both developed below. As I view it, it is not a suggestive analogy but a structural claim about the architecture of cognitive grounding.

For practical purposes, PRU also operates at the level of individual words, tracking how the same word-form functions in two modes:

• x^o: an object-word — a word-token functioning in coordination with a specific experiential constellation. "Dog^o" as uttered while encountering a particular dog. x^o labels {X} — it is the linguistic element integrated into the experiential pattern through encounter.

• x^c: a concept-word — a word-token functioning within linguistic practice. "Dog^c" as it appears in "Dogs are mammals." x^c operates within ⟨x⟩ — the autonomous space where words connect to words.

• $\mathcal{R}(x^o, x^c)$ Reference at the word level — the concept-word coordinating with the object-word, which anchors to {X}.

A child who has acquired only dog^o can point and say "dog!" but cannot answer "what is a dog?" except by ostension. A child with dog^c can answer: "an animal that barks." The concept-word operates in the autonomous linguistic space; the object-word is tethered to encounter.

As I see it, the sequence that builds this structure is:

$\{X\} \rightarrow \{X, x^o\} \rightarrow \{X, x^o, x^c\} \rightarrow \mathcal{R}(x^o, x^c)$

Stage 1 is basic grounding: the pre-linguistic experiential pattern. Stage 2 integrates a label through encounter-events. Stage 3 achieves dual function: the same word can operate in object-mode or concept-mode. Stage 4 is reference proper: the coordination between modes[5].

This sequence is irreversible. It must begin with experiential encounter and proceed through label-integration before achieving the dual-mode structure that makes reference possible. In contrast, LLMs begin and remain at Stage 3 — concept-word-mode mastery without the developmental history that builds it from encounter. They excel at $\mathcal{R}(x^c, y^c)$, coordinating concepts with concepts, because their training corpus is precisely a record of such coordinations. But no amount of concept-concept mastery produces the object-mode anchoring that grounds reference, because x^o requires a history of encounter that text cannot transmit.

3. Returning to Mollo and Millière

3.1. Reconstruction

We can return to Mollo and Millière's teleosemantic argument. Mollo and Millière draw on teleosemantic theories of representational content – (Shea, 2018), (Millikan, 2017), (Neander, 2017) - to specify two conditions for referential grounding:

1. Causal-informational[6]: Internal states must correlate with worldly features through causal chains.

[5] The first two stages are separated here for clarity but available evidence for how infants acquire language suggests that these are tightly intertwined.

[6] The phrase "causal-informational relations" appears twenty-five times in Mollo and Millière's original paper. I adopt their terminology throughout, but note that in several instances what is described as "causation" is better characterized as correlation — as when the authors illustrate the concept: "smoke is *correlated* with fire because it is *caused* by it."

2. Historical/selectional: Those states must have been selected to carry that information — must have the function of tracking worldly features. They argue LLMs satisfy these conditions through three pathways:

Post-training: RLHF (Reinforcement Learning from Human Feedback) introduces extra-linguistic norms (factuality, helpfulness). Human raters reward world-tracking outputs, establishing a selection history with extra-linguistic success conditions.

Pre-training: Next-token prediction implicitly selects for internal states that track worldly regularities. Mechanistic interpretability evidence (Othello-GPT, probing studies) shows LLMs develop representations that model extra-linguistic structure.

In-context learning: Mollo and Millière further argue that mesa-optimization during inference can establish transient world-involving functions without parameter updates. Since this case is even more clearly intra-linguistic than the other two, the critique that follows applies *a fortiori*.

Their conclusion: LLMs achieve referential grounding. Multimodality and embodiment are neither necessary nor sufficient. The argument is sophisticated and represents, in my view, the strongest available case for LLM grounding, however, it does not succeed. The reasons converge on a single structural point: the argument is conducted entirely within ⟨x⟩ — the collective, publicly articulable register that is precisely what the grounding problem says is insufficient.

3.2. Teleosemantics Embeds but Does Not Constitute Grounding

3.2.1. Where the theory earns its force

A frog sits at the edge of a pond. A fly crosses its visual field; retinal ganglion cells fire; the tongue strikes. The frog eats. A frog whose detector

---

The slide is not accidental; it reflects a deeper difficulty with importing teleosemantic vocabulary into LLM analysis, which I address in section 3.4.

misfires — at shadows, at BB pellets, at nothing — goes hungry. Over evolutionary time, only frogs whose detectors reliably tracked flies in their environment survived and reproduced.

This is the scenario that gives teleosemantics its explanatory force. The theory says: the frog's retinal state has the function of indicating prey because ancestors whose states performed that function were differentially selected. This yields normativity without mystery — the state is *supposed to* track flies, and when it fires at a BB pellet it *misrepresents*, because its proper function is fly-detection regardless of what caused the particular firing (Neander, 1991). Normativity and misrepresentation are naturalized in one move[7].

Tree rings carry information about age, yet nobody thinks tree rings have content. What distinguishes the frog is that the information its states carry is useful: useful for catching prey, for surviving, for reproducing. The world, so to speak, *trains* the organism, and it trains against the organism's essential functions. Selection that tracked irrelevant information, however faithfully, would be eliminated. The "teleo-" in teleosemantics does not apply to all purposes, but only to purposes constituted by the organism's embodied predicament, where getting things wrong risks death.

[7] One might object that the frog's fly-catching is itself a teleosemantic case: the tongue-strike has the function of catching flies, selected through evolution. We grant this — indeed, the main text uses this description. But the description operates from the theorist's ⟨x⟩ register: it attributes content, conditions of satisfaction, and normativity to the frog's states as a third-person redescription. Whether the frog's cognitive economy itself requires content-bearing states is precisely what is at issue between teleosemantic and radical enactivist accounts. Our point is that what makes the frog grounded — its dynamic embodied coupling to the fly through sensory transduction, motor readiness, and evolutionary history — is prior to and independent of the teleosemantic redescription. The redescription is accurate as far as it goes, but it presupposes the coupling rather than explaining it. See (Hutto and Myin, 2013, chs. 3–4) for a sustained argument that basic cognition is contentless, and (Thompson, 2010) for the life-mind continuity that makes basic grounding a feature of all organisms, not only linguistically sophisticated ones.

3.2.2. Where it was forged

Teleosemantics earned its credibility entirely on pre-verbal systems: frogs, bees, magnetotactic bacteria (Dretske, 1988), (Millikan, 1989)— organisms whose states are said to acquire content through direct environmental coupling, prior to and independent of language. The theory works for these cases precisely because selection under environmental pressure operates in a register independent of the register in which we *describe* environmenal coupling. That independence is what gives the theory its naturalistic force: (what teleosemantics identifies as) content is explained without presupposing language, meaning, or interpretation.

Human language was forged in the same crucible. Language evolved and develops in the context of coordinated action: shared warning, tool use, teaching, negotiation — activities where linguistic performance is answerable to worldly coping. When a child learns "hot," the selection pressure is not linguistic — it is the burn. The norms governing human language use are downstream of grounded functions. Language, for humans, inherits its normativity from the embodied predicament it serves.

3.2.3. The application to LLMs

It helps to take the framing just introduced — the world *trains* the organism — and run the analogy in the other direction. If LLM training is relevantly similar to what the frog learns about its environment, the comparison should survive reversal. When we reverse it, however, the disanalogy becomes stark.

Mollo and Millière argue that LLMs satisfy the teleosemantic conditions: their internal states stand in causal-informational relations to worldly features (mediated by training data), and those states were selected (through training) to track those features. We do not dispute that there is genuine selection in LLM training. Gradient descent differentially

retains internal configurations; RLHF further shapes outputs against human-evaluated norms. Shea (2018) has argued that the teleosemantic framework extends beyond natural selection to include learning-based selection, which makes the application to LLMs prima facie plausible. The question is: *selection against what*?

For the frog, the world trains against grounded functions — the organism's need to eat, flee, mate, and navigate. The selection criterion and the represented domain are the same: the world. For the LLM, training selects against linguistic adequacy. Next-token prediction retains states that model textual regularities. RLHF retains outputs that human raters approve of — and yes, raters apply worldly norms like factuality, but they apply them as linguistic judgments about linguistic outputs. The utility criterion that gives teleosemantics its explanatory force — utility for an organism coping with its environment — has been replaced by utility for producing contextually appropriate language.

Human raters are grounded, and their judgments channel worldly norms into the training signal. So, the LLM's selection history *borrows* its world-connection from the raters rather than constituting its own. In biological teleosemantics, this separation never arises: the organism whose states acquire content is the same organism whose survival is a selection pressure. The system that represents and the system that is selected are coupled to the same world through the same body. For LLMs, the system being selected and the system whose grounding constitutes the selection norm are different systems entirely. The LLM is selected to match the *outputs* of a process that was grounded, without itself being grounded.

3.2.4. The scope condition

What emerges is a structural mismatch between teleosemantics and its application to LLMs that operates at three levels. The *selection medium* is language ⟨x⟩ rather than the world: the LLM is trained on text, not on environmental encounters. The *normative source* is borrowed: the

world-connection in the training signal derives from the raters grounding, not from the model's own coupling to what it represents. And the utility criterion is linguistic rather than existential: the function the LLM is selected to perform is language production, not worldly coping.

For organisms, all three levels align automatically: the selection medium is the world, the normative source is the organism's own embodied history, and the utility criterion is survival in its environment. That alignment was so seamless that it never needed stating — it was an invisible precondition of the theory. LLMs are the first systems where the three come apart, and their divergence exposes a scope condition that was always present but never visible: teleosemantics requires that selection occur in a medium that makes contact with the organism whose states carry information about it. LLMs make this condition visible by being the first systems whose selection medium is the linguistic residue of worldly contact rather than worldly contact itself.

Teleosemantics, then, presupposes an already-coupled organism and redescribes that coupling in normative terms. Applied to a system that lacks the coupling, the description still goes through, because the theory was formulated at a level of abstraction that does not distinguish direct environmental coupling from statistical inheritance through text. But the grounding does not come with the description. Teleosemantics works well for systems that are already grounded. It is structurally silent about what makes grounding possible in the first place.

3.3. The isomorphism retreat

Mollo and Millière invoke teleosemantics precisely to go beyond mere structural correspondence — they grant, citing Shea, that isomorphism is too cheap. But under indirect mediation, their conditions reduce to isomorphism plus inherited selection history. So, the question becomes whether structural correspondence, even when causally produced, suffices for grounding.

Interpretability studies show that LLM internal states mirror worldly organisation — Othello-GPT develops broad-state representations, probing studies find vectors tracking spatial relations, colour properties, categorical structure. A reader might concede everything in the previous section and still hold that this correspondence, by itself, suffices for grounding.

I argue it does not. A photograph of a cat preserves the structural relations of the scene with extraordinary fidelity: spatial layout, colour relations, relative sizes, occlusion patterns. The causal chain is impeccable — light from the real cat struck the sensor, was transduced, was stored. The photograph satisfies the causal-informational condition and was produced by a device designed to track worldly features faithfully. Yet no one would say the photograph *refers* to the cat. It is a trace — an imprint that preserves structure without contacting what it preserves. An AI-generated image of a cat, visually indistinguishable from the photograph, makes the point sharper: identical structural correspondence, no causal contact with any cat at all. If isomorphism were sufficient for grounding, the generated image would be grounded in a cat that does not exist.

LLMs are in the position of the generated image, not the photograph. Their internal states mirror worldly structure because they were trained on text produced by grounded speakers — they inherit the structural trace that worldly contact impressed upon language. The correspondence is real, non-accidental, and functionally sustained. But correspondence is not contact. A map drawn from other maps preserves the geography without visiting the terrain.

3.4. What causal chains transmit

Mollo and Millière argue that training data carries causal-informational traces of the world into LLM weights: text about fire was written by people who encountered fire; statistical regularities in that text preserve worldly structure; training encodes that structure into internal states.

I grant all of this. The causal chain is real. What needs examination is what the chain transmits and what kind of causal relation it establishes.

When a grounded speaker writes about fire, she performs a specific operation: she converts her experiential engagement with fire — {FIRE}, the multi-modal, affect-laden, *locus*-bound pattern formed through her fire-encounters — into a linguistic artifact. Those sentences encode ⟨fire⟩: the distributional profile of "fire," its co-occurrences, its inferential relations, its sentential behaviour. This is what language does: it retains the collectively shareable structure and discards the irreducibly first-personal. The causal chain from world to LLM thus passes through a conversion — from {X} to ⟨x⟩ — that strips out the experiential pole. What reaches training data is ⟨fire⟩: the collective precipitate of fire-encounters, not the encounters themselves. The LLM, trained on this data, develops internal states that model the structure of ⟨fire⟩ with impressive fidelity. That structure reflects worldly regularities, because ⟨fire⟩ was produced by grounded beings whose language use was shaped by their fire-encounters.

There is a further difficulty. Teleosemantics characterizes systems whose internal states are causally *coupled* to environmental features through sensory transduction. When the fly crosses the frog's visual field, the fly's presence causally modulates the detector's firing in real-time: remove the fly, the firing stops. This is a detection relation — the internal state covaries with the worldly feature because the feature controls the state. When the LLM acquires internal states that mirror the use of "fire" in the training corpus, it experienced no fire. Worldly structure is present in its weights because it was present in the text — which was produced by grounded speakers whose fire-encounters shaped their linguistic output. The causal ancestry is real: fire-encounters are among the distal causes of the LLM's weights. But causal ancestry is not causal coupling. The LLM's internal states track regularities in *text*, not regularities in *the world that produced the text*. To apply "causal-informational" equally to the frog's fly-detection and the LLM's text-trained states is to treat inherited distal causation as if it were the proximal coupling teleosemantics

describes. That is a substantial philosophical commitment, not a terminological convenience.

A concrete scenario sharpens the distinction. Consider a group of humans settling on Mars, carrying with them an Earth-trained LLM. From the first day, the settlers' grounded functions are under new selection pressure. Gravity is 0.38g — "heavy" and "light" begin to shift their experiential anchoring and, with it, their metaphorical extensions. "Outside" no longer means open air; it means lethal vacuum. The respiratory vocabulary — "breathe," "fresh air," "suffocating" — reorganises around the omnipresent dependence on life support. Weather words lose their old referents and acquire new ones as dust storms replace rain. None of this is deliberate linguistic reform. It is language doing what language has always done: reshaping itself under pressure from the embodied predicament of its speakers. The settlers' language updates because their grounding updates — they are coupled to a world, and when the world changes, the coupling pulls language with it.

The LLM need not update. Its internal geometry encodes the distributional profile of "heavy," "outside," "breathe," and "storm" as shaped by the entire history of Earth-bound human text. It will continue generating "step outside for fresh air" as a suggestion for relaxation. It will associate "light" with ease and "heavy" with burden at Earth-calibrated magnitudes. It will produce weather forecasts structured around precipitation. Not because it tracks the settlers' world, but because it tracks text produced by organisms who were tracking a different world. Only when new text — written by settlers whose encounters with Martian conditions have reshaped their language — enters a future training corpus would the LLM's associations begin to shift. Its states do not covary with the world. They covary with text about a world. When the world changes and the text hasn't yet, the LLM is revealed for what it is: a map of Earth carried to Mars, accurate in structure, connected to nothing underfoot.

3.5. A consequence

One might object that RLHF escapes the ⟨x⟩ register. Human raters reward factuality, penalize hallucination — and "is this factually accurate?" seems to be a question about world-correspondence, not linguistic coherence. But examine what actually flows through the training loop. A rater reads an output — a linguistic artifact — and evaluates it against her own beliefs, which are grounded in her {X} but expressed as a judgment about ⟨x⟩. What enters the model is a scalar reward signal that adjusts the probability of producing similar ⟨x⟩-level outputs. RLHF introduces extra-linguistic *norms on text*. It does not introduce extra-linguistic contact with the world. The grounding lives in the rater. What crosses into the model is a judgment about ⟨x⟩ quality, not about {X} itself. The selection pressure is real, but it selects for text that satisfies grounded evaluators.

Mollo and Milliere's framework cannot, even in principle, distinguish between two systems: one that tracks worldly features through experiential contact — $\mathcal{R}(\{X\}, \langle x \rangle)$ fully realized — and one that tracks the linguistic traces of worldly features with enough fidelity to satisfy any ⟨x⟩-level evaluation. This is not an epistemic limitation — not a matter of needing better interpretability tools or more sophisticated probing studies. It is a methodological consequence of conducting the evaluation in the same register as the system under evaluation.

4. Developmental Irreversibility

Recall the sequence introduced in the previous section:

$$\{X\} \rightarrow \{X, x^o\} \rightarrow \{X, x^o, x^c\} \rightarrow \mathcal{R}(x^o, x^c)$$

Stage 1: Basic grounding. The pre-linguistic experiential pattern. Animals have this. The cat encounters mice; {MOUSE} forms through serial encounters. No language involved. This is the ground floor that the standard debate skips.

Stage 2: Label integration. The child encounters dogs in contexts where "dog" is heard. The word enters the experiential constellation as an embodied element, producing {DOG, dog°}. The label does not represent the experience; it is absorbed into it through co-occurrence in encounter-events.

Stage 3: Dual function. Through increased linguistic practice, the same word acquires concept-mode operation: "dog" can now function within language autonomously — "Dogs are mammals," "Is that a dog?" — as well as in object-mode, pointing at particular encounters. The child has both dog° and dog^{c}.

Stage 4: Reference proper. $\mathcal{R}(x^{o}, x^{c})$ — the mature speaker coordinates between modes. She can use "dog" to talk about dogs (concept-mode) and to pick out this particular dog in front of her (object-mode, anchored in {DOG}). Reference is the coordination between these modes, not a single link from word to world.

The sequence cannot run backward. Concept-mode mastery (Stage 3) presupposes the availability of object-mode (Stage 2), which presupposes the formation of {X} through encounter-events (Stage 1). Grounding must occur in the learning process itself. It cannot be attributed retrospectively based on functional success at a later stage.

LLMs begin and remain at Stage 3. They master concept-mode with extraordinary sophistication — inferential relations, distributional structure, sentential behaviour. Mollo and Milliere's arguments from post-training and pre-training both claim that selection pressures operating at Stage 3 can establish what requires Stages 1–2. PRU says this is structurally impossible. You cannot start with concept-mode mastery and work backward to the experiential anchor.

Mollo and Millière correctly argue that multimodality is neither necessary nor sufficient for grounding. PRU explains why: what matters is not the modality of input but the *mode of learning*. Serial, *locus*-bound, concern-laden encounter-events generate {X}. Massively parallel training over batches of data — whether textual or multimodal — does not. Adding a camera to a parallel-trained system produces sensory data without the developmental structure that turns data into experiential patterns.

LLMs fail to achieve referential grounding not because they lack structured coordinations, but because their coordinations remain strictly $\mathcal{R}(x^c, y^c)$. Teleosemantic selection mediated by text cannot recover the x^o term required for reference proper.

5. Empirical evidence: object-mode vs. concept-mode

The {X} / ⟨x⟩ distinction makes an empirical prediction: if grounded speakers use words differently in object-mode (x^o, anchored in {X}) and concept-mode (x^c, operating within ⟨x⟩), and if these different modes of use leave distributional traces, then training data produced by grounded speakers should encode this difference — and LLMs, trained on that data, should preserve it in their internal geometry.

This prediction is confirmed. Analysis of word-usage manifolds in LLM embedding space reveals a characteristic geometric signature. Sentences using a word in experientially grounded contexts — ostensive, perceptual, deictic (x^o usage) — occupy tighter, more clustered manifold regions. Sentences using the same word in conceptually elaborated contexts — abstract, inferential, generic (x^c usage) — occupy more diffuse, distributed regions. The tight cluster is the "rod"; the diffuse spread is the "caps." The rod-and-caps geometry appears across word categories and across models.

The finding might initially look like evidence *for* Mollo and Millière: LLMs have internal structure that reflects the distinction between experiential and conceptual engagement with the world. But consider what it actually shows. The geometric signature exists because grounded speakers use words differently depending on whether they are in encounter-mode or inference-mode. When a grounded speaker writes "the dog jumped onto the couch" (x^o mode), her word choices, syntax, and co-occurrence patterns differ systematically from when she writes "dogs are social animals" (x^c mode). These distributional differences — generated by the coupling R({X}, ⟨x⟩) in the speaker — leave traces in the

training data. The LLM, trained on this data, encodes the traces faithfully. The signature is a collective residue of subjective groundings, preserved in the ⟨x⟩ patterns that those groundings generated.

The LLM thus preserves even the distinction between experientially anchored and linguistically autonomous usage — it has the map and the map's legend marking which regions came from direct survey — without having done any surveying itself.

6. Conclusion

How do we ground "grounding"? Mollo and Millière's disambiguation of five notions of grounding provided the entry point, and their teleosemantic argument for LLM grounding provided the strongest available case to engage with. Dialectical engagement revealed, I hope, something about the problem itself.

Natural language can produce concepts about grounding but not grounding itself. Correspondence is not contact and causal ancestry is not causal coupling. This is because basic grounding runs from world to organism, not from language to world; that language coordinates with a connection already in place, which I explored via the {X}/⟨x⟩/$\mathcal{R}$ notation.

Teleosemantics earned its credibility on pre-verbal organisms whose selection medium was the world, whose normative source was their own embodied history, and whose utility criterion was existential. LLMs are the first systems where they come apart: the selection medium is text, the normative source is borrowed from human raters, and the utility criterion is linguistic adequacy. This is why Mollo and Millière's argument does not succeed.

Attempting to settle a decade-long debate, I hope to have shown that the grounding debate needs tools that can mark the boundary between what language can transmit and what it cannot — and the discipline to respect that boundary when evaluating systems that operate entirely within the transmissible register (⟨x⟩). I argued PRU is one candidate for such a tool.

As I articulated it, the sequence from basic grounding through label-integration to dual-mode reference is developmentally irreversible. And the rod-and-caps finding provided preliminary empirical evidence that the {X} / ⟨x⟩ distinction leaves measurable geometric traces in embedding spaces, which the LLM faithfully preserves without instantiating the distinction that produced them.

Mollo and Millière ask how vector embeddings can acquire meaning. The question presupposes that meaning is the kind of property an internal item can come to possess. On the Wittgensteinian view developed here, this is already the wrong grammar of the problem. Meaning is not first attached to an item and then connected outward to the world. It is constituted in practice: in language-games, in patterns of use, and in the ongoing coupling between speakers and the world they inhabit.

This is why the Mars settlers language updates. Their words shift because their form of life shifts. "Outside," "air," "weather," "heavy," and "dangerous" are pulled into new patterns by the embodied predicament of the speakers. The practice moves, and language moves with it.

LLMs do not participate in such a practice. They encode the linguistic residue of practices carried out by others and manipulate that residue with remarkable competence. Their internal states may preserve world-traces, including traces of how grounded speakers use words under different conditions. But preserving the residue of a practice is not the same as being a participant in the practice that constitutes meaning.

Mollo and Millière's teleosemantic conditions identify when such residue is well tuned to the world from which it descends. They show, at most, that LLMs inherit structured world-traces through text and training. What they cannot identify is what they were never designed to identify: a system whose language is answerable to the world through its own ongoing form of life.

References

Cojocariu, F. (2026) *Pattern-Recognition Unity (PRU): A Framework Specification; A Meta-Language for Grounding, Reference, and the Experience–Language Interface*. Available at http://dx.doi.org/10.2139/ssrn.6285878

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, Mass: MIT Press.

Dumitru, M. (2004). Denotare şi descripţie: un criteriu al referinţei pentru termenii singulari. In *Explorări logico-filozofice* (pp. 50-121). Humanitas.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, Mass.: MIT Press.

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford University Press.

Millikan, R. G. (1989). Biosemantics. *Journal of Philosophy*, 86(6), 281–297. https://doi.org/10.2307/2027123

Mollo, D. C., & Millière, R. (2025). The vector grounding problem. *arXiv preprint arXiv:2304.01481*. https://doi.org/10.48550/arXiv.2304.01481.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge, Mass.: MIT Press.Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of science*, *58*(2), 168–184. https://doi.org/10.1086/289610.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.