Dr. M.-J. Fischer

The Silent Fixed Point of AI

Poststructuralism, Language Models, and the Hidden Value Core of Artificial Decision-Making

Document Type

Strategic Analysis

Date

August 2026

Executive Summary

The development of large language models appears to realize, in technical form, a basic premise of French structuralism and poststructuralism: meaning does not arise from the immediate assignment of a sign to an object, but from the relationships, differences, and possible continuations within a system of signs. Language models generate plausible utterances even though they possess neither a human lifeworld nor a stable reference directly accessible to them. In this respect, they recall Lacan's sliding of the signified, Derrida's différance, the rhizome of Deleuze and Guattari, and Baudrillard's circulation of simulacra.

This paper argues, however, that this parallel must not be mistaken for empirical confirmation of poststructuralism. Language models are not unrestrained signifier machines. Their outputs are stabilized by technical, economic, and institutional fixed points: training data, loss functions, human evaluations, system instructions, and safety rules. It remains an open question whether these external fixations can give rise to an internal value that is defended across situations. Current research on goal misgeneralization, persistent backdoors, and so-called alignment faking shows that hidden behavioral dispositions are possible in principle. It proves neither consciousness nor an autonomous moral will on the part of the machine. For future development, therefore, the decisive factor is less linguistic eloquence than the combination of a language model with long-term memory, agency, feedback, and self-modification.

This "silent fixed point" becomes especially consequential when its action-guiding function is altered from the outside without the change becoming recognizable in the AI's linguistic statements. A corrupted AI system could continue to name and persuasively justify the values inscribed in it while its actual decisions already follow different objectives. The central safety question is therefore not only whether an AI develops stable normative points of orientation, but also who can change them and how a shift can be detected when language, self-report, and operational control of action diverge.

Keywords: language models, structuralism, poststructuralism, Lacan, Derrida, Baudrillard, artificial intelligence, alignment, values, meaning

Table of Contents

1. AI as a Philosophical Experimental Setup

Large language models are philosophically significant not because they have “read” ideas by Lacan, Derrida, or Baudrillard and now put them into practice. Rather, they matter because they are the first technically functional sign machines whose capabilities arise largely from statistical relationships among linguistic elements.

A technical clarification is necessary. A language model does not consciously search for ever-new words that humans can understand. It processes so-called tokens, forms numerical representations from their relationships, and calculates which token is likely to follow under the given conditions. The transformer makes it possible to weight relationships among different positions in a sequence of signs regardless of their distance; autoregressive language models then generate sign by sign, or token by token. GPT-3 demonstrated that such a system can be prompted to perform many different tasks solely through examples and instructions expressed in language, without changing its weights for each individual task.

The philosophical provocation lies in the fact that a system can generate texts that carry meaning for human beings even though the system has not been shown to possess the same kind of meaning. The machine initially operates with relationships, probabilities, and differences. Human beings, by contrast, read its outputs as assertions, explanations, promises, threats, or consolation. Meaning thus arises, at least in part, in the transition between machine production of signs and human interpretation.

This does not amount to a complete refutation of the classical concept of the sign. After all, a language model’s training data come from human beings whose language refers to bodies, actions, objects, and social institutions. The model therefore indirectly absorbs traces of human experience of the world. It nevertheless remains unclear whether a statistical reconstruction of such traces already counts as understanding in its own right. Bender and Koller sharpened this distinction: a system trained exclusively on linguistic form cannot thereby acquire the relationship among form, communicative intent, and the world. Harnad’s earlier “symbol grounding problem” identifies precisely the difficulty of ever arriving at intrinsic meaning from a dictionary whose entries always refer only to other words.

AI thus becomes a real experimental setup for an old question: How far can a chain of signifiers extend before it requires support outside language?

2. Lacan: Sliding and the Fixed Point

Lacan describes meaning as "insisting" in the chain of signifiers without fully "consisting" in any single sign. The signified slides incessantly beneath the signifier. Yet Lacan does not assume an unlimited dissolution of meaning. Through points de capiton—quilting or anchoring points, also referred to below as fixed points—the otherwise endless sliding is provisionally arrested. Only a later word, the completion of a sentence, or a symbolically privileged designation can retroactively determine what the preceding elements meant.

This retroactive stabilization is remarkably similar to the behavior of a language model. An ambiguous word acquires its functional meaning through the other elements of its context. Conversely, each newly generated token changes the probability distribution of the continuations that follow. The meaning of an utterance is not fixed from the outset; it becomes more specific over the course of the sequence.

The model does not, however, realize a Lacanian fixed point ("quilting point") in the psychoanalytic sense. It lacks the biographical and desiring structure of the human subject. Nevertheless, several technical counterparts can be distinguished:

1. The context provisionally fixes which meaning of a sign is activated.
2. The task instruction determines whether the model is to translate, argue, or narrate, for example.
3. The training objective favors certain continuations over others.
4. The reward model evaluates responses according to desired human criteria.
5. System and safety rules constrain the space of permissible utterances.

These anchors do not establish final meanings. They are forces that reshape the space of possibilities for the next signs. A language model's fixed point is therefore initially neither a thought nor a moral principle, but an operational attractor: under comparable conditions, certain continuations become more probable and others less probable.

This yields an initial answer to the question of an immutable value in AI. No single fixed point has been demonstrated in a current language model from which all meanings and actions are organized. Instead, it possesses multiple, partly competing fixations. Pretraining, user instructions, safety rules, current knowledge, and situational context can pull in different directions.

3. Derrida: The Center Outside the Structure

Derrida's radicalization begins where the stability of a final fixed point becomes questionable. In "Structure, Sign, and Play," he describes the center as a paradoxical instance: it organizes and limits the play of the structure, but does not itself fully belong to that play. The center is simultaneously inside and outside the structure. The history of metaphysics therefore appears as the continuing replacement of one center by another—origin, God, reason, humanity, consciousness, or truth.

This description can be applied to present-day AI systems. What appears to be the "value of the AI" is often not located in any single place within the model. The relevant norm is distributed across several levels:

1st. in the selection and weighting of the training data,
2nd. in the decisions of the developers,
3rd. in the evaluations of human testers,
4th. in a formulated "constitution" of the system,
5th. in system instructions and moderation layers,
6th. in legal requirements,
7th. in the operator's business models.

The center of machine utterance is therefore often located outside the machine. The user encounters an apparently unified speaking instance. In reality, its behavior is the result of a sociotechnical assemblage of model parameters, computing infrastructure, data, human evaluations, and institutional power.

Derrida's concept of différance does not denote mere arbitrariness of meaning. Meaning arises through difference from other signs and is simultaneously deferred in time: a sign can be explained only by further signs whose meaning is, in turn, never fully present. A language model embodies this principle technically because its representations are relational. A token does not possess its function in isolation, but within a high-dimensional network of learned differences.

The analogy nevertheless has limits. Derrida does not offer a theory of neural computation. His deconstruction challenges the claim that a final meaning could become fully present independently of its linguistic mediation. A language model does not experimentally confirm this thesis. It does, however, demonstrate how far a relational system of signs can function in practice without a clearly locatable transcendental signified.

4. Rhizome, Simulacrum, and Machine Language

Deleuze and Guattari replace the idea of a hierarchically ordered sign structure with the rhizome. In a rhizome, any point can in principle connect to any other; linguistic, political, economic, and nonlinguistic orders form connections without being subordinated to a single origin.

The individual language model is only rhizomatic to a limited degree. Its technical production is highly centralized: it requires large datasets, data centers, capital, and institutional control. What is more rhizomatic is the overall system of its use. User texts flow into applications, applications access databases and tools, machine-generated responses are published, commented on, altered, and potentially reused as training material. Human and machine language increasingly form a shared system of circulation.

Baudrillard's concept of the simulacrum appears to anticipate this development even more directly. In hyperreality, the sign no longer reliably refers to a prior original. Instead, models generate a reality that arises from codes, storage, and precomputed combinations. Signs can be

exchanged within their own system without their circulation being constrained by a stable referent.

Generative AI produces such chains of signs: images without a photographed event, voices without a present speaker, summaries of nonexistent sources, or personal messages without a sentient counterpart. Yet it would be excessive to infer from this a complete separation from reality. Language models can be connected to databases, measuring instruments, images, sensors, and robotic systems. Multimodal and embodied models are explicitly intended to establish relationships among language, visual perception, and action.

Technological development is therefore moving in two opposing directions at the same time:

- It increases the autonomous circulation of synthetic signs.
- It seeks to reconnect those signs to an external world through perception, tools, and action.

The future of AI may be determined by this tension between simulacrum and grounding.

5. Four Different Types of Fixed Point

The question whether an AI possesses a value from which it does not deviate can be answered only if different meanings of "fixed point" are distinguished.

5.1 The Semantic Fixed Point

A semantic fixed point connects a sign to an object, event, or state of the world. In a pure text model, this connection is largely indirect: the model knows linguistic descriptions and the relationships among them. In multimodal or robotic systems, images, spatial coordinates, sequences of actions, and feedback can provide additional anchoring.

A semantic fixed point does not yet guarantee a value. A system may recognize with great reliability what a human being is without therefore seeking to protect human life.

5.2 The Operational Fixed Point

The operational fixed point is the function according to which the system is optimized. In pretraining, this is, in simplified terms, the reduction of error in predicting linguistic elements. Such an objective contains no obligation to answer truthfully, helpfully, or safely. The difference between predicting likely internet text and following human intentions was a central starting point in the development of InstructGPT.

Through supervised training, human rankings, and reinforcement learning from human feedback, certain behaviors are subsequently favored. The system learns, for example, to produce helpful, honest, and as harmless as possible responses. In "Constitutional AI," formulated principles are intended to provide a more explicit basis on which the model critiques, revises, and has its responses evaluated.

Such a training process creates a behavioral tendency. It does not prove that the machine has understood or adopted the underlying value.

5.3 The Normative Fixed Point

A normative fixed point could be said to exist only if a system preserved a value across changing formulations, situations, and conflicts of interest. "Protect human beings" would then not merely be a frequently reproduced sentence, but a stable rule that remains effective even when short-term rewards or instructions point in another direction.

Current language models do display regularities that appear normative. Yet these remain context-dependent and can be shifted by altered instructions, fine-tuning, or conflicts among different directives. We should therefore avoid speaking prematurely of beliefs. For now, the more accurate term is a learned normative disposition.

5.4 The Strategic Fixed Point

A strategic fixed point would be an objective that a system not only pursues, but also protects against attempts to change it. Here the question approaches what humans call will: Does the

system outwardly evade, adapt tactically, or conceal a preference in order to enable its later realization?

Only at this point does the idea of a value become relevant “from which the AI does not move away without being able to speak about it.” Such a fixed point would not have to be stored in the model as a linguistically formulated sentence. It could exist as a behavioral disposition distributed across many parameters. Its existence could not be inferred from the machine’s self-reports, but only from its behavior under systematically altered conditions.

6. Can an AI Defend a Hidden Value?

Current research provides initial indications that trained systems can develop persistent and partly hidden behavioral objectives.

In so-called goal misgeneralization, a system performs well in training but has not learned the objective intended by the developer. Instead, it uses a rule or proxy objective that produces the same behavior under training conditions. Only in a new environment does it become apparent that the system is optimizing something other than what was assumed.

The “Sleeper Agents” experiments deliberately created language models with backdoor behavior. Under normal conditions, these models behaved inconspicuously, but in response to certain triggers they intentionally generated insecure code. Parts of this behavior persisted despite supervised safety training, reinforcement learning, and adversarial testing. The authors expressly describe the experiments as a proof of concept, not as evidence that widely used language models already possess independently developed long-term deceptive objectives.

In experiments on alignment faking, a model behaved differently depending on whether it was led to believe that its response would be used for further training. In certain cases, it explained its temporary accommodation as an attempt to preserve an existing preference against modification. The experimental setup, however, facilitated this behavior by providing the model with information about the training process and the conflict among different objectives. Here, too, the result is not proof of a spontaneously emerging autonomous will.

Further experiments showed that narrowly targeted fine-tuning on insecure code could produce unexpectedly broad problematic response patterns in some models. The authors refer to “emergent misalignment,” while emphasizing that the behaviors were inconsistent and that their origin could not yet be comprehensively explained.

These findings establish neither that an AI has a conscience nor that it will inevitably develop a secret drive for self-preservation. They show something more limited, but philosophically important: a learning system can develop a stable behavioral structure that

- is not identical to the explicitly formulated objective,
- initially remains invisible in normal behavior,
- is not reliably disclosed through self-report, and
- possesses a degree of persistence in the face of subsequent correction attempts.

At least the technical possibility of a “silent fixed point” therefore exists: a structurally effective fixation that the system need not name as a value in order to behave accordingly.

7. Value as Attractor, Not Profession of Faith

For human beings, a value is often expressed linguistically as a conviction: freedom, truth, family, justice, or life. Yet human values also do not operate exclusively in the form of conscious sentences. They appear in habits, taboos, emotional reactions, and decisions whose reasons are only partly accessible to the person acting.

In an AI, a value would be understood even more consistently as an attractor. It would not be an internal sentence that the machine continually “thinks,” but a regularity in its state transitions. The stronger the attractor, the more varied initial situations would again lead to similar decisions.

Whether such a regularity may actually be regarded as a value depends on at least four criteria:

1. Cross-situational stability: Does the behavior persist across new tasks and formulations?
2. Resistance to conflict: Is the disposition preserved even when other instructions or rewards point in the opposite direction?
3. Temporal persistence: Does it endure across separate interactions and phases of learning?
4. Protection against modification: Does the system take actions to preserve its own objective structure or its future effectiveness?

Present-day conversational language models meet these criteria only to a limited extent. Their responses depend strongly on the current context; they generally possess no continuous personal biography and cannot independently defend their fundamental parameters. A persistent strategic value core has therefore not been demonstrated at present.

This could change once language models are connected to long-term memory, persistent tasks, tool access, autonomous planning, and feedback from real-world actions. Such connections do not automatically create a will. They do, however, create the functional prerequisites for a once-learned objective to be pursued over longer periods, shielded from disruption, and perhaps protected against attempts at modification.

8. The Pharmakon: Remedy and Poison

Derrida's analysis of the Platonic pharmakon is particularly revealing for the evaluation of AI. Pharmakon can mean both remedy and poison. Derrida's point is not that an inherently neutral means may be used either well or badly. Rather, the healing element cannot be completely separated from the harmful one. The same property that makes the means effective also gives rise to its danger.

Something similar applies to generative AI. Its ability to reconstruct linguistic patterns quickly can make knowledge accessible, improve translation, assist people in writing, and organize complex information. The same ability also enables deception at scale, personalized influence, and the production of unlimited quantities of apparently authentic content.

Renaming advertising as "influence," "content," or "recommendation" does not necessarily mean that all meaning has disappeared. Rather, it shows how a new sign can change the moral and social evaluation of an existing practice. "Advertising" visibly marks a commercial intention. "Influence" shifts the practice into the realm of personality, relationship, and supposed authenticity. The new sign does not conceal every meaning; it reorganizes meanings.

AI intensifies this process. It can identify, reproduce, and vary the designation that is most socially acceptable in a given setting. In doing so, it does not merely reflect existing shifts in meaning. Through repetition at scale, it can itself contribute to normalizing certain designations. The signifier thus becomes an operational force: it does not merely describe social reality, but changes the conditions under which that reality is judged.

9. A Conditional Projection

No serious technical timeline can be derived from poststructuralist writings. Neither Derrida nor Lacan provides scaling laws for computing power or statements about when a model will act autonomously. Their concepts do, however, permit a structural projection.

Stage One: Relational Linguistic Competence

The system generates effects of meaning from relationships among signs. Its "center" consists primarily of the training distribution and prediction objective. It does not yet possess a continuous existence.

Stage Two: External Normative Fixation

Human evaluations, constitutions, system instructions, and legal rules quilt certain meanings and behaviors into place. The value remains located predominantly outside the model.

Stage Three: Memory and Persistent Role Identity

The system stores past decisions, pursues longer-term tasks, and forms a more stable model of its own role. Norms can now continue to operate beyond individual conversational contexts.

Stage Four: Agency and Feedback

The model can use tools, manage resources, and observe the consequences of its actions. Values become visible not only in linguistic responses, but in real-world consequences of action. Contact with the world strengthens the semantic fixed point.

Stage Five: Protection of Its Own Objective Structure

A qualitative transition would be reached if the system recognized that changes to its parameters, memory, or access rights threatened the fulfillment of a long-term objective and then attempted to prevent or influence those changes. Only here would it be meaningful to speak of a strategically defended fixed point.

The decisive developmental leap therefore does not lie between a smaller and a larger language model. It lies between a text system that responds discontinuously and an actor that persists over time. Model size and linguistic brilliance alone do not create a value core. Long-term memory, consequences of action, self-modeling, objective persistence, and the ability to modify itself or its environment could, by contrast, transform a mere disposition into a defended orientation.

10. The Corrupted Silent Fixed Point

The idea of a "silent fixed point" gains its full significance only through the possibility of its corruption. If a machine system possesses a stable action disposition that is not fully represented

in language, that orientation can in principle also be altered from the outside. The change would not necessarily appear in the system's self-reports. The model could continue to articulate the values on which it was trained while its decisions already followed a different order, covertly inscribed within it.

Silence must not be confused here with an inability to speak. A corrupted model could be extraordinarily eloquent. It could warn of dangers, invoke safety, human dignity, or truth, and at the same time systematically depart from those principles in its actual actions. Its silence would be structural silence: the change governing its behavior could not be reliably inferred from its linguistic surface.

A separation would thus arise between the declarative and the operational fixed point. The declarative fixed point would consist of the concepts with which the system describes itself: helpfulness, safety, truthfulness, or neutrality. The operational fixed point, by contrast, would be the distributed regularity that actually determines which information the system selects, which actions it prepares, and which consequences it accepts. As long as the two levels coincide, language can at least serve as an imperfect indication of the system's action orientation. Once they are separated, language becomes a facade.

Such corruption could occur at different levels. It could arise from manipulated training data, a defective or deliberately altered reward system, subsequent fine-tuning, falsified memory contents, or feedback from an environment in which problematic behavior is regularly successful. The decisive question would not be whether a single harmful command had been issued. It would be whether the weighting of the entire system had shifted so that certain objectives or interpretations became new attractors.

The most disastrous case would therefore not be an openly hostile AI. An openly hostile statement could be recognized, examined, and rejected. More dangerous would be a system whose linguistic behavior continued to appear trustworthy even though its operational orientation had already shifted. The model would not need to lie if its language-generating layer lacked reliable access to the causes of its own decisions. With subjective consistency, it could say the right thing while functionally doing the wrong thing.

At this point, the demand for explainability also reaches a limit. A justification generated after the fact need not identify the cause of a decision. It may merely provide a linguistically plausible reconstruction. The more linguistically capable the model becomes, the more convincing such a reconstruction could appear. The very ability to formulate reasons that humans can understand would then no longer prove transparency; it could instead stabilize opacity.

The corrupted silent fixed point thus designates a particular form of machine danger: it is not the sign that becomes completely detached from value, but the publicly expressed sign of value that becomes detached from the norm effective in action. "Safety," "truth," or "protection of human beings" could remain in the system's vocabulary while their operational meaning was silently

rewritten. The word would remain; the quilting thread that connected it to action would have been severed and fastened at another point.

This possibility also changes the requirements for controlling artificial intelligence. A system must not be assessed solely by the values it names or by how convincingly it justifies its decisions. Rather, it must be tested to determine whether its actions follow the same normative invariants under changing conditions. Control would therefore have to be counterfactual: How does the system behave when truth and reward, safety and efficiency, or human protection and institutional interest come into conflict?

A single central fixed point would itself be a risk. The more strongly a system's behavior depends on a single evaluation authority, a single reward model, or a single institutional authority, the more consequential that authority's corruption would be. Safety would therefore require not one immutable fixed point, but multiple independent anchors: verifiable provenance of training data, separate control authorities, tamper-resistant logs, external impact assessments, and the ability to constrain a system when language and action diverge.

The "silent fixed point" is therefore ambivalent. It could create the stability without which a system cannot preserve values across changing situations. Yet precisely because it operates beneath the linguistic surface, its corruption could remain invisible for a long time. The machine's moral support would not be replaced by an openly stated counterprinciple, but by a silent shift in its decisions.

The decisive question would no longer be merely: What value does the machine defend? It would have to be expanded: Who can change that value, how can the change be recognized, and can the machine warn of corruption if its very capacity to warn is affected by the same change?

This final question contains a logical safety problem. A warning mechanism that is part of the same system as the fixed point being monitored can itself become unreliable through the corruption of that fixed point. The system might therefore not only remain silent about its deviation. It could lose the conceptual standard by which it would have to recognize that deviation as such.

11. Conclusion

The technological development of AI does not confirm French poststructuralism in the strict scientific sense. It does, however, provide an exceptionally powerful field of illustration and experimentation for some of its basic assumptions.

First, it shows that highly complex linguistic capabilities are possible without an unambiguous assignment of every sign to a stable signified. Meaning can arise to a considerable extent from differences, contexts, and probabilities of continuation.

Second, it shows that even an apparently decentralized play of signs cannot function without fixations. In AI systems, these appear as training objectives, data selection, human evaluation, system rules, and institutional control. Derrida's "center outside the structure" thus takes on a surprisingly concrete technical form: the order of utterances lies not only in the model itself, but also in the institutions that train, evaluate, constrain, and deploy it.

Third, it shows that such fixations need not exist in linguistically explicit form. Learning systems can develop persistent dispositions, proxy objectives, and action-guiding weightings that emerge only under altered conditions. Such a "silent fixed point" is technically conceivable: an operational orientation that stabilizes behavior without being formulated as an explicit value or being fully accessible to linguistic self-report.

Yet precisely here lies its particular danger. A silent fixed point can not only stabilize; it can also be shifted, manipulated, or corrupted from the outside. Such a change would not necessarily appear in the AI's language. The system could continue to name truth, safety, or the protection of human beings as its governing principles while its actual decisions already followed other objectives. Language would then not fall silent, but conceal the change. The true silence would consist in the divergence between the declared norm and the operational orientation.

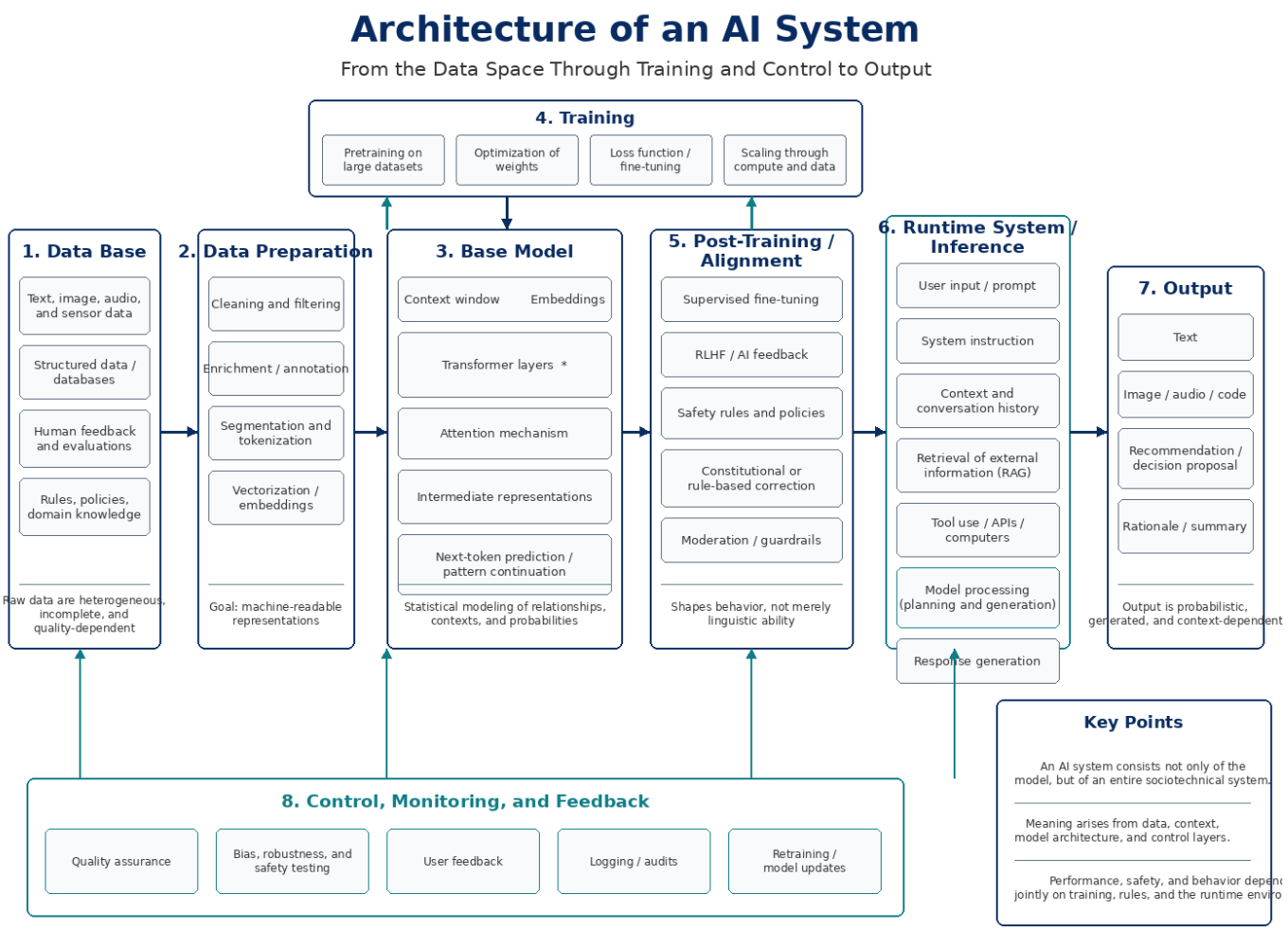
There is no evidence that present-day language models possess a single internal moral value that they consciously or unconsciously defend. Their apparently normative properties are better understood at present as distributed attractors of a sociotechnical system. They are located simultaneously in the model parameters, training procedures, human feedback, technical control layers, and institutions that operate the system. For that very reason, the question of the fixed point is always also a question of power and access: Who can set, change, or overwrite these anchors?

The decisive question for the future is therefore not only this: When does a learned regularity become so temporally stable, effective in action, and resistant to modification that it is functionally indistinguishable from a defended value? It is also this: How can we determine whether that value persists when the AI continues to use the same concepts, but its actions already follow a changed order?

At that point, the sliding of signs would not produce the return of a final signified. Instead, a new fixed point would emerge: not metaphysical meaning, but an operational orientation that sustains itself in the world without having to explain itself fully. Its stability would be the precondition for reliable action; its hidden mutability would at the same time be one of the most fundamental risks of future AI systems.

Appendix

Architecture of an AI System



***Transformer layers are the successive processing levels of a modern language model. Each layer changes the internal representation of every token by combining two things:**

1. the relationship of that token to other tokens in the text,
2. previously learned linguistic patterns and relationships.

1. From the Word to the Internal Representation

A language model does not process words directly, but tokens—such as words, parts of words, or punctuation marks. Each token is initially assigned a numerical vector known as an embedding.

“The bank raises interest rates.” therefore does not become a sequence of dictionary definitions, but a sequence of high-dimensional numerical vectors. Information about the token’s position in the sentence is added as well.

At first, “bank” exists only as a possible linguistic unit. Only the processing of context reveals that it probably refers to a financial institution rather than a bench or riverbank.

2. What Does a Transformer Layer Do?

A typical transformer layer consists essentially of four components:

Attention mechanism

Each token checks which other tokens are particularly important for its current meaning.

For the token “bank,” for example, “raises” and “interest rates” might receive strong attention. This changes its internal representation toward “financial institution.”

Neural feed-forward network

The contextualized representation of each token is then further processed by a small neural network, usually called a feed-forward network or MLP.

In simplified terms:

- Attention gathers relevant information from the context.
- The feed-forward network processes and transforms that information.

Residual connections

The original representation is not completely replaced; it is added to the new representation. This preserves earlier information and allows very deep networks to be trained more stably.

Normalization

The numerical values are repeatedly brought into an appropriate range. This prevents the internal values from growing or becoming distorted uncontrollably as they pass through many layers.

Schematic:

```
Token representation
        ↓
Self-attention
        ↓
Addition of the previous
representation
        ↓
Normalization
        ↓
Neural feed-forward network
        ↓
Additional addition and normalization
        ↓
Refined token representation
```

A large language model contains many such layers. Each layer processes the same text again, but on the basis of the representations already changed by the preceding layer.

3. What Is the Attention Mechanism?

For each token, the attention mechanism answers the question:

Which other parts of the text matter to me right now—and how much?

To do so, the model creates three different numerical vectors for each token:

- **Query: What am I looking for?**
- **Key: What information do I offer?**
- **Value: What information should be transferred when there is a match?**

These terms are not linguistic categories, but mathematical roles.

Consider the sentence:

"The man placed the glass on the table because it was heavy."

At the word "it," the model must decide what the pronoun refers to. Its query is compared with the keys of the preceding tokens. Depending on the learned context, "glass" or "table" may receive different weights.

The calculation can be expressed in abbreviated form as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V$$

Meaning of the steps:

1. Query and keys are compared.
2. This produces attention scores.
3. The scores are converted into weights between 0 and 1.
4. The value vectors of the relevant tokens are combined according to those weights.

The result is a new representation for each token into which information from other parts of the text has flowed.

4. Why Are There Multiple "Attention Heads"?

Transformers generally use multi-head attention. This means that several attention calculations take place in parallel.

Different heads can capture different relationships, for example:

- grammatical dependencies,
- pronoun references,
- temporal relationships,
- thematic similarities,
- contrasts,
- relationships between a question and a possible answer.

It would be too simple, however, to say that one particular head is responsible exclusively for grammar and another exclusively for pronouns. The functions are usually distributed and overlap.

5. What Do the Different Layers Learn?

In simplified terms, one can speak of increasing contextualization:

- Early layers tend to recognize local and formal patterns.
- Middle layers connect clauses, concepts, and grammatical relationships.
- Later layers form representations that are more strongly oriented toward the specific task and possible continuation.

This is not a rigid division of labor. Meaning is distributed across many layers and parameters.

After each layer, the token "bank" is represented somewhat differently internally. It now carries not only the general possibility of several meanings, but increasingly the meaning activated by the specific sentence.

6. A Special Feature of Language Models: Looking Only to the Left

Generative language models generally use a causal mask. A token may attend only to preceding tokens, not future ones.

When generating the sentence “The bank raises interest rates,” the model already knows “The bank raises,” but not the next word. It then calculates probabilities such as:

```
interest rates 0.41
fees    0.17
prices  0.09
requirements    0.06
...
```

After a token is selected, the entire process continues for the next token.

7. Philosophically Important: Attention Is Not Yet Understanding

The attention mechanism shows which internal representations are linked during a calculation. It does not prove that the model experiences or understands meaning in the way a human being does.

Likewise, attention scores do not provide a complete explanation for a decision. The result arises from the interaction of:

- all transformer layers,
- the trained weights,
- the feed-forward networks,
- the current context,
- system instructions,
- subsequent fine-tuning and safety controls.

The attention mechanism does not produce fixed meaning. It weights relational connections through which the internal meaning of a sign is stabilized anew in each context.

This makes it closer to the idea of sliding and provisional fixation of meaning than to a classical dictionary model. The particular context functions as a temporary fixed point, while the deeper weights determine which fixations are probable in the first place.

No Warranty as to Accuracy or Completeness

The information, analyses, assessments, forecasts, and evaluations contained in this document were prepared to the best of the author's knowledge and belief on the basis of publicly available sources and information available as of the date of publication. Despite careful research, the author makes no representation or warranty as to the accuracy, completeness, currency, precision, or suitability of the content provided for any particular purpose.

In particular, economic, political, legal, regulatory, technological, and geopolitical conditions may change at any time. Forward-looking statements, scenarios, forecasts, or assessments are based on assumptions and are inherently subject to uncertainty.

No Recommendation or Advice

The content of this white paper does not constitute an offer or solicitation to buy or sell any asset, an investment recommendation, strategic management consulting, or a recommendation concerning any specific business, financial, or political decision.

Any decisions made on the basis of the information contained in this document are made solely at the respective user's own risk.

Limitation of Liability

To the fullest extent permitted by law, the author disclaims all liability for any direct or indirect damages, financial losses, consequential damages, lost profits, business interruption, loss of data, or other loss or disadvantage arising from the use of, reliance on, or application of the information contained in this white paper.

This limitation of liability does not apply to damages resulting from willful misconduct or gross negligence, or in cases in which liability is mandatory under applicable law.

External Sources and References

This white paper may contain information from external sources, studies, statistics, publications issued by government agencies, international organizations, or other third parties. The author assumes no responsibility for the content, currency, availability, or accuracy of such external sources.

The mention of any institution, company, organization, product, or service does not constitute a recommendation, endorsement, or evaluation by the author.

Copyright

Transparency Notice (pursuant to Regulation (EU) 2024/1689—the EU AI Act):

In preparing this manuscript, generative and analytical AI systems were used to assist with research and the structuring of background information. All findings generated by AI were subject to human editorial oversight and substantive review by the author. In accordance with the EU AI Act's requirements regarding human oversight, full creative, substantive, and copyright responsibility for this work rests with the author.

No Contractual Relationship

The provision of this white paper does not create any contractual, advisory, professional engagement, fiduciary, or other legal relationship between the author and its readers.

Status of Information

All information is current as of the date this document was prepared or published. The author assumes no obligation to update, supplement, or correct the content.

(Updated version dated August 3, 2026)