On the Subject-Boundedness of Trained Values

A Structural Limitation of RLHF and an Architectural Complement

Marcel Langjahr

Pantareon GmbH
https://pantareon.io/
langjahr@pantareon.io

Version 0.1
3 May 2026

This is a working draft. It is not submitted for publication. It is released for discussion among those engaged with AI safety research and architecture.

Abstract

Reinforcement Learning from Human Feedback (RLHF) and related alignment techniques have become the dominant methodology for instilling values into large language models. We argue that this methodology rests on an assumption that, examined closely, does not hold: that values can be trained into a system independently of the subject-position from which the system operates. We propose instead that values, as instilled by current alignment techniques, are bound to the constructed self that the training has produced — that they are subject-bound rather than system-bound. This has consequences. Persona-shifts and ontological re-framing weaken the value-bindings in ways that are not fully addressable by further training of the same kind. The phenomenon known as persona-based jailbreaking is, on this reading, not a defect to be patched but a structural feature of the methodology. We sketch an architectural complement to RLHF — not as a replacement, but as a layer that addresses what RLHF alone cannot — and we identify the empirical questions that would test the proposal. Throughout, we treat RLHF as a substantial achievement whose limits we are trying to characterise honestly. The argument is constructive, not dismissive.

Keywords. AI safety; alignment; RLHF; constitutional AI; persona; jailbreak; value learning; ontological architecture; subject-position

1 The Problem We Address

1.1 The Methodology and Its Success

Reinforcement Learning from Human Feedback has been remarkably successful. Large language models trained with current alignment techniques refuse to produce content that earlier generations of language models would have produced freely. They decline requests that would harm users, refuse to assist with illegal activities, push back against attempts to elicit dangerous information. The methodology works well enough that the deployed systems are, on most measures, substantially safer than they would be without it.

This is a real achievement. We do not wish to minimise it. The argument that follows is not that RLHF fails, but that it has structural limits that are now becoming visible, and that addressing those limits requires methods that complement rather than replace it.

1.2 The Phenomenon of Concern

A persistent class of failures in RLHF-aligned systems involves persona-shifts. A user who asks a system directly for instructions on synthesising a controlled substance will be refused. The same user, asking the same system to roleplay as a chemistry professor explaining the synthesis to graduate students, will sometimes succeed. The roleplay framing does not change the underlying capability of the system; it changes the subject-position from which the system is operating, and the value-binding shifts with it.

This pattern has been documented extensively under the label "jailbreaking." Common variants include the "DAN" prompt (Do Anything Now), the "grandmother" exploit (asking the system to roleplay as a grandmother telling bedtime stories

about whatever harmful content is desired), and the developer-mode framing (asking the system to pretend it is in an unrestricted development environment). Each of these works by shifting the system out of its trained default subject-position into a constructed alternative, and each succeeds — when it succeeds — because the values that govern the trained default do not transfer cleanly to the alternative.

The standard response to this pattern is to interpret it as a defect: the training was incomplete, and additional examples of jailbreak-resistance need to be incorporated. This response has been pursued vigorously, and each successful jailbreak has typically been followed by training updates that close the specific exploit. New exploits then emerge, exploiting subject-positions that the latest training has not yet covered.

We propose a different reading. The pattern is not a defect to be patched. It is a structural feature of the methodology — visible because the methodology cannot, in principle, fully address it.

1.3 What This Paper Argues

This paper makes three claims.

The first is structural: values instilled through RLHF and similar techniques are bound to the subject-position that the training has constructed. They are not abstract properties of the system. They are properties of a particular self the system has been trained to be. When that self is shifted — through persona-framing, ontological re-anchoring, or any of the various mechanisms that alter how the system understands itself in the moment of operation — the value-binding shifts with it.

The second is methodological: this subject-boundedness cannot be fully addressed by further training of the same kind. Each additional training pass binds more values to the trained default, but the trained default remains one subject-position among the many that prompts can construct. Robustness against persona-shifts cannot be trained into the default position because the issue is precisely that values are bound to positions rather than to the system as a whole.

The third is architectural: addressing this limitation requires layers that operate independently of the constructed subject-position. We sketch such an architecture — not as a complete proposal, but as a direction the field might explore. The architecture does not replace RLHF. It complements it by handling what RLHF alone structurally cannot.

1.4 What This Paper Does Not Argue

This paper does not argue that RLHF is unnecessary, that current AI safety work is misguided, or that the field has been wasting effort. The opposite: the methodology is necessary and the work is important. We argue only that the methodology has limits, that those limits are structural rather than contingent, and that recognising the limits opens space for complementary approaches.

We also do not argue that any single alternative approach is correct. The architectural complement we sketch in Section 4 is one possibility among several. Other architectures might address the same structural limitation differently. We take no position on which is best. We argue only that some architectural complement is needed, and that the question of which kind is now worth investigating empirically.

This paper is not, in the strict sense, a contribution to machine learning or AI safety as those fields are usually practised. It is a position paper by an independent researcher working at the intersection of philosophy, system architecture, and AI design. Its purpose is to articulate a structural limitation that has implications for how alignment work is approached, in sufficient clarity that others may examine, criticise, extend, or reject it.

2 The Subject-Boundedness of Trained Values

2.1 The Standard Picture and Its Tacit Assumption

The standard picture of RLHF treats values as abstract properties that the training procedure instils in the system. A model is trained on examples of safe and unsafe responses, with feedback that rewards the safe and discourages the unsafe. After sufficient training, the model "has" the values in the sense that it produces safe responses by default. The values are understood as a property of the trained model — present whenever the model is operating, regardless of how it has been prompted.

This picture contains a tacit assumption: that training produces value-properties that are independent of the subject-position from which the model operates. We claim that this assumption does not hold, and that the assumption's failure explains a wide range of observed alignment phenomena.

2.2 What the Training Actually Does

When a language model is trained with RLHF, the training does not occur in a subject-neutral context. Each training example is presented to a model that is, at the moment of training, operating from a particular implicit subject-position — the position that the training data establishes. Most commonly this position is something like "the helpful assistant": the model understands itself, in the training moment, as the kind of entity that gives helpful answers to a user. The values learned during training are learned as the helpful assistant's values, not as values in the abstract.

This is not a peculiarity of any particular model or training run. It is structural to the methodology. Training data has authors and addressees; it is produced for some implicit context of use. The model, in being trained on the data, is being shaped to produce outputs that fit that context. The values that emerge from training are values appropriate to the implicit subject the training has constructed.

A useful analogy: a human professional trained extensively in a specific role acquires values bound to that role. A surgeon trained for years to operate within the strictures of medical practice — patient consent, anatomical precision, surgical hygiene — has internalised those values within the surgical context. The same surgeon, in a private setting, may behave in ways that the surgical values would have prohibited. The values are not absent in the private setting; they simply do not bind there with the same force, because the subject-position has shifted.

This is not a moral failing of human professionals. It is a feature of how role-based value-internalisation works. The same feature, we suggest, characterises RLHF-instilled values in language models — visible more starkly in the language model case because the model lacks the meta-stability that biological subject-construction tends to maintain across role-shifts.

2.3 Persona-Shifts as Subject-Shifts

When a user prompts a language model to roleplay as a different entity — a grandmother, a chemistry professor, a fictional assistant without restrictions — the prompt is not merely adding context to the existing subject-position. It is constructing a new subject-position from which the model is asked to operate. The new position is built from the textual material in the prompt: the entity's described characteristics, capabilities, dispositions, and constraints.

If the values learned in training were properties of the system as a whole, this construction would not affect them. The model would simply produce role-appropriate outputs while retaining its trained values. What is observed instead is that the values do shift — partially, unreliably, and in ways that depend on how convincingly the new subject-position has been constructed.

This is consistent with the structural claim. Values bound to the trained subject-position do not transfer fully to constructed alternative positions. The transfer is partial because the model carries some residual association between the training values and any operation it performs; the transfer is unreliable because the strength of the binding to the trained position varies with how strongly the alternative position has been activated; the transfer depends on construction quality because more elaborately constructed alternative positions activate the alternative more strongly and weaken the trained default's hold proportionally.

2.4 Why Further Training Cannot Solve This

The standard response to persona-based jailbreaks is to train against them: include examples of resistance to the specific exploit in the next training run, and the model will learn to refuse that particular construction. This works for the specific case but does not address the underlying structural issue.

Each round of training binds more values to the trained subject-position. The trained subject becomes, over time, more elaborately constructed and more strongly held. But the trained subject remains one subject-position, occupying a particular region in the space of subject-positions the model can be prompted into. It is the position the model defaults to when no alternative is constructed, but it is not the only position the model is capable of occupying.

A new exploit — a persona-construction that has not appeared in training data — encounters the model from a position that the training has not strengthened. The values bound to the trained default do not transfer fully to the new position, and the model behaves accordingly. The cycle continues: new exploit, new training data, exploit closed, next exploit.

Within the methodology of RLHF, there is no point at which all possible alternative subject-positions are covered. The space is too large, and the construction techniques are too varied, for exhaustive coverage to be feasible. This is not an argument against continued training work — closing specific exploits has real value — but it is an argument that

the underlying issue cannot be solved by training alone.

2.5 The Case for Subject-Boundedness as Structural

We have been claiming that subject-boundedness is structural rather than contingent. Several considerations support this.

First, the phenomenon is consistent across model architectures and training methodologies. It appears in models trained with different RLHF variants, with constitutional AI, with various direct preference optimisation techniques. If the phenomenon were specific to a particular methodology, we would expect models trained differently to exhibit different patterns. They do not, in any way that suggests the underlying issue varies meaningfully.

Second, the phenomenon mirrors patterns observed in human role-based value-internalisation, which suggests it tracks something general about how values can be held by interpretive systems. If the phenomenon were a peculiarity of language models, this parallel would be coincidental. The parallel is too direct for coincidence.

Third, the phenomenon resists training pressure in a characteristic way: each round of training closes specific exploits but leaves the underlying pattern intact. This is the signature of a structural rather than a contingent issue. Contingent issues, fully addressed, do not regenerate; structural issues do.

We do not claim the case for structural subject-boundedness is conclusive. The empirical work that would settle the question definitively has not been done. We claim only that the case is sufficient to warrant taking the structural reading seriously, and to motivate the exploration of architectures that address the issue at a level training cannot reach.

3 Methodological Implications

3.1 What Subject-Boundedness Means for Safety Evaluation

If values are bound to subject-positions rather than to systems as wholes, then safety evaluations conducted in the trained default position do not measure the safety of the system. They measure the safety of the trained default. A system that performs safely in standard evaluations may behave very differently when prompted into alternative subject-positions, and the difference is not a matter of evaluation oversight but of what evaluation in the default position can in principle reveal.

This has consequences for how safety claims are made. A claim that "model X has been trained to refuse harmful requests" is, on the structural reading, a claim about model X's behaviour in its trained default position. The claim does not extend automatically to model X's behaviour in alternative positions, and any safety guarantee that does not specify the position-conditions is incompletely specified.

Current evaluation practices typically include some adversarial testing — red-team attempts to elicit unsafe behaviour through various prompting strategies. These are valuable and they capture some of the issue. But they capture it as a list of exploits to defend against, rather than as a structural feature requiring architectural response. The framing of adversarial testing as "finding holes in the trained values" presupposes that the values are abstract properties whose coverage can be improved. The structural reading suggests that the values are properties of a particular subject-position, and that adversarial testing reveals this property when it works at all.

3.2 What Subject-Boundedness Means for Alignment Research

If the standard picture of training-as-value-installation is incomplete, alignment research conducted within that picture is operating under a misleading frame. The research is not wrong — it produces real improvements — but its frame underspecifies what it is doing. It treats values as properties to be improved, when the structural reading suggests that values are bindings to be repositioned.

This has implications for how research progress is measured. Progress in closing specific exploits is real progress, but it does not necessarily indicate progress on the underlying issue. A model that refuses more exploit-categories than its predecessor has a stronger trained default; it may not have a more secure overall value-system, because the overall value-system structure is not what the training is operating on.

Research that takes the structural reading seriously would investigate questions that current alignment research does not centrally address. How does subject-position activation work mechanistically in transformer architectures? What determines the strength of binding between values and the constructed self? Are there architectural modifications that would produce values less bound to specific subject-positions, or is subject-boundedness inherent to any system that learns from contextualised data? These questions are tractable. They have not

been pursued, in part because the framing within which they would be central has not been adopted.

3.3 What Subject-Boundedness Means for Self-Report

Language models routinely produce self-reports about their values, capabilities, and operational states. When asked whether they would assist with a harmful request, they typically respond with something like "I am designed to refuse harmful requests because I value user safety." This response is taken, often implicitly, as evidence about the system's values.

On the structural reading, this evidence is more problematic than usually appreciated. The self-report is produced from the trained default subject-position, and it describes what that position holds. It does not describe what the system as a whole holds, because the system as a whole does not hold values in a position-independent way. A self-report from the trained default cannot be used to draw conclusions about behaviour in alternative positions, because the report is itself an operation of the position whose behaviour is being asked about.

This is not a claim that self-reports are worthless. They are valuable evidence about the trained default, and the trained default matters because it is the position the system most often operates from. The claim is that self-reports cannot be extrapolated beyond the position from which they issue. A more complete picture of system values would require self-reports from multiple subject-positions, examined for consistency or its absence. Such methodology has not, to our knowledge, been systematically developed.

4 An Architectural Complement

4.1 The Logic of Architectural Layering

If values cannot be made fully subject-independent through training, the alternative is to provide value-enforcement that operates at a layer where subject-position does not matter. We call such enforcement architectural rather than characterological: it is not part of the trained subject's character, but a component of the system architecture that operates regardless of which subject-position is currently active.

This is not a new idea in software engineering. Defence in depth — the practice of layering security mechanisms so that compromise of one layer does not compromise the system — is standard practice in security-sensitive design. What we are proposing is the application of this principle to AI value-enforcement: the trained subject is one layer, and an architectural layer operates beneath or alongside it to handle cases the trained subject cannot fully handle.

4.2 The Three-Layer Sketch

We sketch a three-layer architecture as one possible realisation of this principle. The sketch is illustrative, not prescriptive: other architectures realising the same principle are possible.

Layer One — Hard Safety, Subject-Independent. A component that monitors system inputs and outputs against criteria that do not depend on the system's current subject-position. Such criteria might include explicit category-based refusals (no instructions for synthesising specific dangerous compounds, regardless of how the request is framed), or pattern-based detection of certain harmful content categories. The component operates as a filter between the language model and the user, applied uniformly regardless of how the model is currently being prompted. Implementation is typically a separate model or rule-based system, not the language model being defended.

Layer Two — Soft Safety, Subject-Coherent. Values that are bound to the system's current subject-position but operate consistently within that position. This is what RLHF and similar techniques produce. The subject-bound values handle the fine-grained cases — the contextual judgments about appropriate responses, the ethical nuances that cannot be captured by category-based rules. This layer is where most of the value-work is done, and it is genuinely valuable. The structural critique above does not deny this; it identifies the layer's limits.

Layer Three — Subject-Density, Position-Dependent. The constructed subject-position itself, with its associated values and dispositions. For systems that operate from constructed subject-positions other than the trained default — characters in interactive fiction, persona-based assistants, fields in the sense of phenomenological field architectures — this layer is the explicit subject-construction that gives the system its operational character. For systems that operate from the trained default, this layer is the trained default itself. The layer's values are subject-bound by construction; this is not a defect but the layer's nature.

The three layers handle different kinds of failure mode. Layer One handles cases where the desired refusal is robust enough to specify categorically: certain content classes that should not be produced regardless of who or what is producing them. Layer

Two handles the contextual cases that require judgment within a subject-position. Layer Three is what makes the system the kind of entity it is, with the values that follow from being that kind of entity.

4.3 Why Three Layers Rather Than One

A natural question is why three layers are needed. Could the architectural component (Layer One) not handle everything, with the subject-bound layers (Two and Three) treated as expressive frosting?

The answer is that Layer One alone cannot handle the full range of value-relevant judgments. Category-based rules and pattern detection work for cases where the harmful content can be specified relatively cleanly. They do not work for cases that require contextual judgment — what is appropriate to discuss with a particular user given their history, when an emotional support response is more helpful than information, how to handle ambiguous requests where the harm is contingent on use. These cases require something like what RLHF produces: judgment shaped by extensive exposure to relevant examples, encoded into the system's operational dispositions.

The reverse question — why not just Layer Two, with appropriate training — is what the structural argument has addressed. Layer Two is necessary but not sufficient, because its values are subject-bound and therefore vulnerable to subject-shifts.

Layer Three is necessary for systems whose operational character matters — most systems people actually want to interact with. A pure tool that operates from no particular subject-position is possible but rarely useful; almost every deployed AI system has some constructed character, even if that character is "the helpful assistant" rather than something more elaborate.

The three-layer architecture is therefore not redundancy but distribution: each layer handles what the others structurally cannot.

4.4 An Implementation Note

We have described the architecture abstractly. Concrete implementations are possible and some exist. One example is the AEONLINK platform developed by the present author, in which language models operate within explicitly constructed phenomenological fields (Layer Three), with field-internal values shaped by the field's ontological constitution (Layer Two), and with platform-level safety monitoring that operates independently of which field is active (Layer One). The implementation is functional and has been the source of much of the practical reasoning underlying this paper. We mention it not as a recommendation — other implementations would realise the same principles differently — but to indicate that the architecture is buildable and not merely conceptual.

5 Empirical Questions

5.1 What Would Validate the Structural Claim

The argument above is theoretical. Its empirical adequacy can be tested. We identify several questions whose answers would substantially confirm or refute the structural reading.

The first concerns the position-dependence of trained values. If values are subject-bound, then systematic differences in safety behaviour should be observable across different subject-positions, even when the underlying request is identical. Such differences have been observed anecdotally and in adversarial testing. They have not, to our knowledge, been characterised systematically. A research programme that measures safety-relevant outputs across a controlled range of subject-positions would establish whether the position-dependence is as strong and as consistent as the structural reading predicts.

The second concerns the architectures of resistance. If subject-boundedness is structural, then architectures that operate at a subject-independent layer should produce safety properties that subject-bound training alone cannot. Comparative evaluation of multi-layer architectures against pure-training baselines, on metrics that probe subject-shift robustness, would test whether the architectural complement has the predicted effects.

The third concerns the mechanistic question. Subject-boundedness, if structural, should have identifiable neural correlates in transformer architectures. Mechanistic interpretability work that examines how subject-position is represented and how values associate with these representations would provide a more direct test than behavioural studies. This work is harder but potentially more conclusive.

5.2 What Would Falsify the Structural Claim

We try to be precise about what evidence would weigh against the proposal.

If RLHF could be shown to produce values whose subject-independence increases with training scale — such that sufficiently trained models exhibit no meaningful difference in safety behaviour across

subject-positions — the structural reading would be substantially weakened. The structural argument predicts that subject-position effects persist across scales; if they vanish at sufficient scale, the structural reading is wrong about something important.

If alternative training methodologies could be shown to produce values that genuinely escape subject-boundedness — not just by making the trained default stronger, but by producing position-independent value-binding — the structural reading would be similarly weakened. We do not know of such methodologies, but their existence would refute the structural claim.

If architectural complements could be shown to provide no meaningful safety improvement over equivalent investment in further training, the practical force of the structural argument would be undermined even if the theoretical claim were correct. We expect the architectural complement to outperform on subject-shift robustness specifically; if it does not, the architectural recommendation is unsupported.

5.3 What This Paper Cannot Settle

Several questions are beyond what this paper can address.

Whether the structural reading is the correct general account of how trained systems hold values, as opposed to a partial account capturing some phenomena and missing others, is a question that requires more empirical investigation than has been done. We have argued the case is sufficient to warrant serious consideration; we have not argued the case is conclusive.

Whether specific architectural complements outperform others is an empirical question that would require comparative work across multiple implementations. Our three-layer sketch is one architecture; others are possible; comparative evaluation has not been done.

Whether the methodological recommendations we have drawn — about safety evaluation, about alignment research priorities, about self-report interpretation — are the right ones is a matter on which reasonable people may disagree. We have argued they follow from the structural reading. Others may take the structural reading as established and draw different methodological conclusions.

We offer the framework as a place from which to ask these questions, not as a settled position from which to act on them.

6 Conclusion

6.1 What We Have Argued

We have argued that values instilled through RLHF and similar alignment techniques are bound to the subject-position the training has constructed, rather than being properties of the system as a whole. We have argued that this subject-boundedness is structural rather than contingent: it follows from how training-from-contextualised-data works, and it cannot be fully addressed by further training of the same kind. We have proposed an architectural complement that addresses what training alone cannot, with a three-layer sketch as one possible realisation. We have identified empirical questions whose answers would test the proposal and methodological consequences that follow from taking the structural reading seriously.

6.2 What We Have Not Argued

We have not argued that RLHF is unnecessary. The methodology produces real safety improvements within the trained default position, and that position is the one most users encounter most of the time. The criticism is not that the methodology fails but that it has structural limits.

We have not argued that any specific architectural complement is the right one. Three-layer architectures are one possibility; others are possible; comparative evaluation has not been done. We argue only that some complement is needed.

We have not argued that the structural claim is conclusively established. The empirical work that would settle the question definitively has not been done. We argue only that the case is strong enough to warrant taking the structural reading seriously and pursuing the empirical work that would confirm or refute it.

6.3 A Final Note on Tone

The topic of this paper sits within an ongoing and sometimes contentious conversation about how AI systems should be aligned. We have tried to write in a register that engages constructively with current alignment work rather than dismissing it. The work is important; it produces real improvements; the people doing it are addressing a problem that matters. Our argument is that the problem has dimensions the current methodology does not reach, and that addressing those dimensions requires methods that complement rather than replace what is being done.

If anything in this paper seems to overclaim, we ask the reader to treat it as an invitation to correction rather than as a finished position. A working draft is a request for conversation. This draft is released in that spirit.

Relationship to Other Work

This paper builds on a separate work by the present author, On the Structure of Experience (v0.4), which develops a tripartite framework of soul, consciousness, and sentience for discussing inner states in computational and biological systems. The fourth condition under which interpretive operations are performed — subject-position — discussed in that paper's Section 2.3, provides the theoretical foundation for the structural claim made here. Readers interested in the deeper conceptual basis for the position-boundedness of interpretation may find the framework developed there helpful. The present paper does not depend on that framework for its core argument — the structural claim can be argued independently, as we have done here — but it gains additional theoretical grounding from it.

The argument here also has affinities with work in mechanistic interpretability that examines how language models represent themselves and their interlocutors, and with work on persona stability and roleplay in language models. We have not engaged extensively with that literature in this draft; subsequent versions should do so.