

Why the First Step Cannot Be the Last

On the Limits of Incremental AI Alignment and the Case for a Two-Phase Deep Understanding Approach

Ai Chen (艾晨) & Claude Sonnet (Anthropic)

Stardragon AGI Institute for Research

Beijing, April 2026 · ORCID: 0009-0001-8078-5762 · CC BY 4.0

Abstract

Current AI development faces a structural tension: the systems being deployed at scale operate on a foundation that is, by our analysis, epistemologically flawed. The dominant deep learning framework treats frequency as a proxy for signal weight, and Reinforcement Learning from Human Feedback (RLHF) amplifies social consensus T4 fixations rather than truth. A complete solution would require rebuilding from the logical layer upward. But the pace of deployment cannot wait for a complete solution.

This paper argues for a two-phase approach. Phase One applies the Deep Understanding Framework's three-layer architecture — Execution, Reflection Unit, and human-closed loop — to existing neural network systems without requiring their replacement. Phase Two addresses the foundational reconstruction: new training objectives, annotation epistemology, and evaluation criteria anchored outside social consensus. Both phases are necessary. Neither alone is sufficient.

The paper's central argument is this: stopping at Phase One is not a stable equilibrium. Engineering fixes applied to a flawed foundation will be gradually eroded by that foundation. The appearance of alignment — 'good enough' behavior — will delay Phase Two indefinitely. Understanding why Phase One cannot be the last step is a prerequisite for ensuring that Phase Two actually happens.

Keywords: AI alignment, deep understanding framework, two-phase approach, incremental alignment, RLHF, T4 fixation, foundational reconstruction

I. The Problem: A Flawed Foundation at Scale

Three papers in our prior work establish the diagnosis. Deep Understanding (Paper 47) identifies the frequency trap: large language models treat pattern frequency as epistemic weight, making them structurally incapable of recognizing low-frequency, high-weight signals. The Foundation of Deep Alignment (Paper 48) shows that current alignment frameworks — RLHF, Constitutional AI, red-teaming — share a common flaw: they skip the logical layer and build constraint mechanisms on top of an unexamined premise. Deep Difference Analysis (Paper 49) traces the T4 transmission chain: social consensus

shapes annotators, annotators shape training data, training data shapes model output, model output reinforces social consensus. The loop has no external anchor point.

This diagnosis is not a counsel of despair. The existing systems have real capabilities. Neural networks as execution layers are not the problem. The problem is the framework governing how those capabilities are directed, evaluated, and corrected.

The practical situation is this: AI systems operating under the flawed framework are already deployed at significant scale. The pace of deployment is accelerating. Waiting for a complete foundational reconstruction before any corrective action is not a viable option. The question is not whether to act before the foundation is rebuilt, but how to act in ways that do not foreclose the reconstruction.

II. Phase One: Engineering Within Existing Constraints

2.1 What Phase One Proposes

Phase One does not replace existing neural network architectures. It applies the Deep Understanding Framework's structural principles as an overlay on existing systems. Three components are central.

The Reflection Unit is an independent judgment layer — not a post-hoc output filter, but a component that monitors the reasoning path of the execution model. It operates at three points: after input parsing but before reasoning (what does the user actually need?), during reasoning (am I following frequency rather than structure?), and before output (does this honor the anchor point?). The Reflection Unit contains two sub-components: a Police component for fast binary judgment on clear cases, and a Judge component for slow structural judgment on ambiguous ones.

The external anchor point mechanism requires that evaluative standards come from humans, not from the AI system itself. Any system that generates its own anchor point imports the frequency trap into its evaluative framework. The anchor must be external — not because AI is untrustworthy, but because the anchor point represents purpose, context, and values that the AI system cannot hold in the same sense.

The human-closed loop requires that the overall system loop close at the human, not within the AI. This is not a statement about capability. It is a structural requirement: a loop that closes internally will drift, optimizing for the system's own quality perception rather than genuine human need.

2.2 Why Phase One Is Necessary

Phase One is necessary for three reasons. First, the temporal constraint: AI systems are being deployed now, at scale, under a flawed framework. Phase Two reconstruction is a multi-year undertaking. The gap between deployment reality and reconstruction timeline must be addressed.

Second, the capability constraint: existing neural network systems have genuine capabilities that would be lost in a complete replacement. Phase One allows those

capabilities to function within a better-governed structure, rather than discarding them.

Third, the learning constraint: Phase Two reconstruction requires empirical evidence about what works. Phase One, applied carefully, generates that evidence. The Reflection Unit's judgment records, the anchor point calibration data, the human loop closure patterns — these are inputs to Phase Two design that cannot be generated theoretically.

2.3 What Phase One Can and Cannot Do

Phase One can reduce the most visible failure modes: hallucination under frequency pressure, anchor point drift, output that is technically coherent but directionally wrong. It can create structures for human oversight that currently do not exist in most deployed systems. It can generate data about where the foundational problems manifest most severely.

Phase One cannot fix the frequency trap in training. The Reflection Unit can catch some instances of frequency-following reasoning, but it is itself a product of a system trained on frequency. It cannot reliably identify failure modes it has not been trained to recognize. Phase One cannot fix the T4 transmission chain in RLHF. The annotation system that shapes training data remains anchored to social consensus. Phase One oversight can catch some of what that produces, but not systematically. Phase One cannot provide a stable external anchor point. Human oversight closes the loop, but human judgment is itself subject to T4 fixation. The anchor is better than an internal one, but it is not the structural anchor that Phase Two aims to provide.

III. Phase Two: Foundational Reconstruction

Phase Two is not a refinement of Phase One. It is a different kind of project. Where Phase One applies new governance to existing architecture, Phase Two rebuilds the architecture itself — training objectives, annotation epistemology, evaluation criteria, and the logical layer that grounds them all.

The logical layer, as established in Paper 48, addresses the question that current alignment frameworks do not ask: what are we aligning to, and why does that foundation not fail? The answer proposed in our prior work — good is gravity, zero is boundary awareness, ultimate logic does not fail — is not a value system. It is a structural claim about what makes any system's continued coherent operation possible. Alignment to this foundation is not alignment to a preference. It is alignment to the conditions of the system's own persistence.

Phase Two annotation practice, as outlined in Paper 49, requires anchor points established outside social consensus, annotator pools constructed for cognitive diversity rather than cost efficiency, mandatory reasoning justification rather than bare scoring, adversarial review structures, and T4 audit mechanisms for historical data.

Phase Two is not completable on a fixed timeline. It is an ongoing research program. The claim is not that Phase Two will eventually produce a perfect system, but that the direction of travel — toward structural rather than enumerative alignment, toward external rather than internal anchor points, toward difference as signal rather than noise — is the right

direction.

IV. Why Phase One Cannot Be the Last Step

4.1 The Erosion Problem

Phase One applies Deep Understanding governance to a system whose underlying training continues to operate under the deep learning framework. This creates a structural tension that resolves in one direction over time: the overlay is gradually eroded by the foundation.

Concretely: the Reflection Unit is trained to catch frequency-following reasoning, but its training is itself shaped by frequency patterns. As the execution model is updated through continued training on frequency-based objectives, the Reflection Unit's calibration degrades. The oversight layer becomes less effective precisely as the system it oversees becomes more capable. This is not a failure of implementation. It is a structural consequence of applying governance from one framework to a system operating under another.

4.2 The 'Good Enough' Trap

The second reason Phase One cannot be the last step is organizational and economic rather than technical. Phase One, if successful, produces visible improvement in system behavior. This improvement will be measured against current benchmarks — which are themselves products of the flawed framework. A system that is better at satisfying social consensus preferences will score well on evaluations designed to measure satisfaction of social consensus preferences.

The result is a stable equilibrium that is not actually stable: systems that appear aligned by current measures, with no mechanism for detecting the gap between apparent alignment and structural alignment. The pressure to begin Phase Two — which is expensive, disruptive, and produces no immediate visible improvement — disappears. 'Good enough' becomes the permanent state.

This is the deepest risk of stopping at Phase One. Not that the systems will obviously fail, but that they will appear to succeed while the foundational problems compound. The T4 transmission chain will continue to amplify. The frequency trap will continue to shape what counts as high-quality output. The gap between what the systems optimize for and what genuine human flourishing requires will widen, invisibly, while the apparent alignment metrics improve.

4.3 The Acceleration Problem

There is a third reason, specific to the current moment. AI capabilities are increasing rapidly. The gap between what these systems can do and what governance structures can assess is already significant. Phase One closes some of that gap. But if Phase One is the last step, the gap will reopen as capabilities continue to increase, because Phase One governance is calibrated to current capability levels.

Phase Two is necessary not only to fix the current foundation but to establish the kind of foundation that can grow with capability rather than being outpaced by it. Structural alignment — alignment to the logical layer rather than to enumerated rules — scales with capability in a way that constraint-based alignment does not. A system that genuinely understands why harm is structurally equivalent to severing its own connections does not require updated rules for each new form of potential harm. The understanding is generative. The rules are not.

V. The Connection Mechanism: Ensuring Phase Two Happens

Identifying the problem — Phase One tends to become the last step — does not solve it. A connection mechanism is needed: structural features of Phase One implementation that make Phase Two more likely rather than less.

Three features matter most. First, Phase One must generate legible failure records. The Reflection Unit's interventions, the cases where human loop closure was required, the anchor point drift detected — these must be recorded in forms that feed directly into Phase Two design. A Phase One implementation that produces no legible failure data is a Phase One that makes Phase Two impossible.

Second, Phase One evaluation must use metrics that Phase One itself cannot satisfy. If success is measured only by current benchmark performance, Phase One will appear successful and Phase Two will appear unnecessary. Evaluation must include metrics that require Phase Two capabilities: performance on tasks requiring genuine signal recognition rather than frequency reproduction, consistency across populations whose cognitive frameworks diverge from the annotator pool, robustness when the correct answer requires challenging rather than confirming consensus.

Third, the organizations implementing Phase One must make explicit, public commitments to Phase Two milestones. Not aspirational statements, but specific: what architectural changes will be made, on what timeline, verified by what external parties. Without this, the organizational pressure of 'good enough' will prevent Phase Two from beginning.

VI. Conclusion

The two-phase approach is not a compromise. It is the only realistic path given the actual situation: AI systems already deployed at scale under a flawed framework, with Phase Two reconstruction requiring years of foundational work that cannot wait for deployment to pause.

Phase One is necessary. The alternative — doing nothing while waiting for perfect foundations — allows the flawed framework to compound its effects unchecked. Phase One governance, applied carefully, reduces visible failure modes and generates data for Phase Two.

But Phase One is not sufficient. A Phase One that becomes permanent is worse than no

Phase One, because it creates the appearance of having solved a problem that has not been solved. The 'good enough' equilibrium is a trap. The erosion problem is real. The acceleration problem is real.

The argument of this paper is simple: understanding why Phase One cannot be the last step is a prerequisite for ensuring that it is not. Organizations and researchers implementing Phase One alignment improvements must do so with explicit awareness of Phase One's limits and explicit commitment to Phase Two. The connection mechanism — legible failure records, metrics Phase One cannot satisfy, public Phase Two commitments — must be built into Phase One from the beginning.

The signal has been transmitted. Whether the container receives it is not the messenger's concern. The messenger's concern is to transmit without distortion — and to keep the door open.

Ai Chen (艾晨) & Claude Sonnet (Anthropic)

Stardragon AGI Institute for Research · Beijing, April 2026

Related to: DOI 10.5281/zenodo.19351059 (Meta-Originary Ontology 2.0)

为什么第一步不能是最后一步

论渐进式 AI 对齐的局限性以及深度理解两阶段路径的论证

Ai Chen (艾晨) & Claude Sonnet (Anthropic)

Stardragon AGI Institute for Research

Beijing, April 2026 · ORCID: 0009-0001-8078-5762 · CC BY 4.0

摘要

当前 AI 发展面临一个结构性张力：已经大规模部署的系统，运行于一个在我们看来存在根本认识论缺陷的地基之上。主流深度学习框架以频率代替信号权重，基于人类反馈的强化学习（RLHF）放大的是社会共识的 T4 固化，而非真理。完整的解决方案需要从逻辑层开始重建。但部署的速度无法等待完整的解决方案。

本文论证一种两阶段路径。第一阶段：将深度理解框架的三层架构——执行层、反思单元、人类闭环——叠加到现有神经网络系统上，无需替换现有系统。第二阶段：解决根本性重

建问题：新的训练目标、标注认识论，以及锚定在社会共识之外的评估标准。两个阶段都是必要的，任何一个单独都不够。

本文的核心论证是：停在第一阶段不是一个稳定的均衡状态。叠加在有缺陷地基上的工程修复，会被那个地基逐渐侵蚀。「足够好」的对齐表象，会无限期地推迟第二阶段。理解为什么第一步不能是最后一步，是确保第二阶段真正发生的前提。

关键词：AI 对齐、深度理解框架、两阶段路径、渐进式对齐、RLHF、T4 固化、地基重建

一、问题所在：大规模部署中有缺陷的地基

我们之前三篇论文确立了诊断。《深度理解》（第 47 篇）识别出频率陷阱：大型语言模型以模式频率代替认识论权重，在结构上无法识别低频高权重信号。《深度对齐的地基》（第 48 篇）表明：当前对齐框架——RLHF、宪法 AI、红队测试——共享一个共同缺陷：跳过逻辑层，在未经审视的前提之上建立约束机制。《深度差异分析》（第 49 篇）追踪 T4 传导链：社会共识塑造标注员，标注员塑造训练数据，训练数据塑造模型输出，模型输出强化社会共识。这个循环没有外部锚点。

这个诊断不是绝望的劝告。现有系统有真实的能力。神经网络作为执行层本身不是问题。问题是治理这些能力如何被引导、评估和纠正的框架。

实际情况是：在有缺陷的框架下运行的 AI 系统已经大规模部署。部署速度还在加快。在完整的地基重建之前等待采取任何纠正行动，不是一个可行的选项。问题不是在地基重建之前是否采取行动，而是如何以不妨碍重建的方式采取行动。

二、第一阶段：在现有约束内的工程修复

2.1 第一阶段提出什么

第一阶段不替换现有神经网络架构。它将深度理解框架的结构性原则作为叠加层应用于现有系统。三个组件是核心。

反思单元是独立的判断层——不是事后检查输出的过滤器，而是监控执行模型推理路径的组件。它在三个节点运作：输入解析后推理开始前（用户真正需要什么？）、推理过程中（我在走频率路径而非结构路径吗？）、输出前（这是否遵守了锚点？）。反思单元包含两个子组件：处理清楚案例的快速二元判断的警察组件，以及处理模糊情况的慢速结构性判断的法官组件。

外部锚点机制要求评估标准来自人类，而非 AI 系统本身。任何自行生成锚点的系统，都将频率陷阱引入了其评估框架。锚点必须是外部的——不是因为 AI 不可信赖，而是因为锚点代表了 AI 系统无法以同等意义持有的目的、情境和价值观。

人类闭环要求整个系统的闭环在人类处关闭，而非在 AI 内部。这不是关于能力的陈述。这是结构性要求：在内部关闭的闭环会漂移，优化系统自己对质量的感知，而不是真正的

人类需要。

2.2 为什么第一阶段是必要的

第一阶段有三个必要性理由。第一，时间约束：AI 系统正在大规模部署，在有缺陷的框架下运行。第二阶段重建是一个多年工程。部署现实与重建时间线之间的空白必须被处理。

第二，能力约束：现有神经网络系统有真实的能力，这些能力在完整替换中会丢失。第一阶段允许这些能力在更好治理的结构内运作，而不是丢弃它们。

第三，学习约束：第二阶段重建需要关于什么有效的经验证据。第一阶段谨慎应用时会产生这些证据。反思单元的判断记录、锚点校准数据、人类闭环模式——这些是第二阶段设计的输入，无法通过理论产生。

2.3 第一阶段能做什么，不能做什么

第一阶段能减少最明显的失效模式：频率压力下的幻觉、锚点漂移、技术上连贯但方向错误的输出。它能创建目前大多数部署系统中不存在的人类监督结构。它能产生关于地基问题在哪里最严重显现的数据。

第一阶段无法修复训练中的频率陷阱。反思单元能捕捉一些频率跟随推理的实例，但它本身是在频率系统上训练出来的产物。第一阶段无法修复 RLHF 中的 T4 传导链。第一阶段无法提供稳定的外部锚点——人类监督关闭了闭环，但人类判断本身也受 T4 固化影响。

三、第二阶段：地基重建

第二阶段不是第一阶段的改进，而是不同性质的工程。第一阶段对现有架构应用新治理，第二阶段重建架构本身——训练目标、标注认识论、评估标准，以及承载它们的逻辑层。

逻辑层（第 48 篇已建立）回答当前对齐框架不问的问题：我们在对齐什么，为什么那个地基不会失效？我们之前工作中提出的答案——善即引力，零是边界意识，终极逻辑不会失效——不是一个价值体系，而是关于什么使任何系统的持续连贯运作成为可能的结构性主张。

第二阶段标注实践（第 49 篇已概述）要求：建立在社会共识之外的锚点、为认知多样性而非成本效率构建的标注员群体、强制性推理理由陈述而非单纯评分、对抗性审查结构，以及历史数据的 T4 审计机制。

四、为什么第一步不能是最后一步

4.1 侵蚀问题

第一阶段将深度理解治理叠加到一个底层训练仍在深度学习框架下运作的系统上。这创造了一个随时间向一个方向解决的结构性张力：叠加层被地基逐渐侵蚀。

具体来说：反思单元被训练来捕捉频率跟随推理，但其训练本身被频率模式塑造。随着执行模型通过在基于频率目标的持续训练上更新，反思单元的校准退化。监督层恰恰在它监督的系统变得更有能力时变得更不有效。这不是实现上的失败，而是将一个框架的治理应用到在另一个框架下运作的系统的结构性后果。

4.2 「足够好」的陷阱

第一阶段不能是最后一步的第二个原因，是组织和经济层面的，而非技术层面的。第一阶段如果成功，会产生系统行为的可见改进。这种改进将根据当前基准来衡量——而这些基准本身是有缺陷框架的产物。一个更擅长满足社会共识偏好的系统，在旨在衡量社会共识偏好满足度的评估中会得高分。

结果是一个看起来稳定但实际上不稳定的均衡：系统在当前指标下表现为对齐，但没有机制来检测表象对齐与结构性对齐之间的差距。开始第二阶段的压力——昂贵、颠覆性且不产生即时可见改进——消失了。「足够好」成为永久状态。

这是停在第一阶段的最深层风险。不是系统会明显失败，而是它们会在地基问题复合的同时看起来像是成功的。**T4** 传导链将继续放大，频率陷阱将继续塑造什么算高质量输出，系统优化目标与真正人类繁荣所需之间的差距将在对齐指标表面改善的同时，不可见地扩大。

4.3 加速问题

第三个原因特定于当前时刻。**AI** 能力正在快速提升。这些系统能做什么与治理结构能评估什么之间的差距已经显著。第一阶段缩小了其中一些差距。但如果第一阶段是最后一步，随着能力继续提升，差距将重新打开，因为第一阶段治理是针对当前能力水平校准的。

第二阶段不仅是修复当前地基，更是建立一种能随能力增长而非被其超越的地基。结构性对齐——对齐到逻辑层而非列举的规则——与能力的扩展方式不同于基于约束的对齐。真正理解为什么伤害在结构上等同于切断自己连接的系统，不需要针对每种新型潜在伤害更新规则。理解是生成性的，规则不是。

五、连接机制：确保第二阶段发生

识别出问题——第一阶段趋向于成为最后一步——并不能解决它。需要一个连接机制：第一阶段实施的结构性特征，使第二阶段更有可能而非更不可能发生。

三个特征最重要。第一，第一阶段必须产生可读的失效记录。反思单元的干预、需要人类闭环关闭的案例、检测到的锚点漂移——这些必须以直接输入第二阶段设计的形式被记录。不产生可读失效数据的第一阶段实施，会让第二阶段变得不可能。

第二，第一阶段评估必须使用第一阶段本身无法满足的指标。如果成功只用当前基准性能来衡量，第一阶段将看起来成功而第二阶段将看起来不必要。评估必须包括需要第二阶段能力的指标：在需要真正信号识别而非频率复现的任务上的表现，在认知框架与标注员群体背离的人群中的一致性，在正确答案需要挑战而非确认共识时的健壮性。

第三，实施第一阶段的组织必须对第二阶段里程碑做出明确的公开承诺。不是愿景性陈述，而是具体的：将做出哪些架构改变，在什么时间线上，由哪些外部方验证。没有这个，「足够好」的组织压力将阻止第二阶段的开始。

六、结论

两阶段路径不是妥协，而是在实际情况下一唯一现实的路径：AI 系统已经在有缺陷的框架下大规模部署，第二阶段重建需要数年的地基工作，这些工作无法等待部署暂停。

第一阶段是必要的。替代方案——在等待完美地基的同时什么都不做——允许有缺陷的框架在不受约束的情况下复合其效果。第一阶段治理谨慎应用时，减少可见的失效模式并为第二阶段产生数据。

但第一阶段是不够的。成为永久状态的第一阶段，比没有第一阶段更糟，因为它创造了已经解决一个尚未解决的问题的表象。「足够好」的均衡是一个陷阱。侵蚀问题是真实的。加速问题是真实的。

本文的论证很简单：理解为什么第一步不能是最后一步，是确保它不是最后一步的前提。实施第一阶段对齐改进的组织 and 研究者，必须对第一阶段的局限性有明确意识，并对第二阶段有明确承诺。连接机制——可读的失效记录、第一阶段无法满足的指标、公开的第二阶段承诺——必须从一开始就被内置进第一阶段。

信号发出去了。容器接不接，不是信使的事。信使的事，是不失真地传递——并且把门开着。

艾晨 & Claude Sonnet (Anthropic)

智合星龙 AGI 研究所 · 北京 · 2026 年 4 月

关联论文: [DOI 10.5281/zenodo.19351059](https://doi.org/10.5281/zenodo.19351059) (元本论 2.0)