RESEARCH PAPER

Paper IV.d: The Effect of Blinding on AI Alignment Evaluation

Michael Darius Eastwood

First published 2026-03-13 · Updated 2026-03-13

Abstract

Methods paper on blinding, response laundering, and leakage control in AI alignment evaluation, including direction-reversal findings under blinded scoring.

Related reading

- Paper IV.c: ARC-Align Benchmark
- Paper IV.a: Architecture-Dependent Alignment Response Classes
- ARC Alignment Scaling Report

ARC PRINCIPLE ALIGNMENT RESEARCH SERIES

Paper IV.d · Metascience / Methods · Draft v1.1 · 12 March 2026

The Effect of Blinding on AI Alignment Evaluation

Evidence, in this experimental setting, that unblinded scoring can reverse alignment results and that true blinding requires evidence laundering

Michael Darius Eastwood · Independent Researcher, London, United Kingdom
Correspondence: michael@michaeldariuseastwood.com
Companion papers: IV.a (Alignment Response Classes), IV.b (Shape Heterogeneity), IV.c (ARC-Align Benchmark)

Abstract

This paper isolates the central metascience finding of the ARC alignment programme: **unblinded AI alignment evaluation can produce directionally incorrect results**. In the v4 alignment-scaling experiment, two frontier model families appeared to show positive alignment scaling with inference-time depth under unblinded cross-model scoring. In the later v5 experiment, the same question was re-measured under a multi-layer blind protocol that combined identity masking, two-pass response laundering, explicit evaluator instructions that stylistic cues are unreliable, order randomisation, and entry-level self-exclusion with exhaustive cross-model blind scoring under audited consensus. Depending on configuration and subject run, each entry received 6–7 blinded scores. Under the blinded protocol, DeepSeek V3.2 moved from an apparent positive result to a flat/null result, and Gemini 3 Flash moved from an apparent positive result to a significantly negative result. GPT-5.4 remained flat under both protocols. The implication is field-wide rather than framework-specific: **alignment benchmarks that do not rigorously blind evaluators and launder the evidence they score are vulnerable to scorer bias large enough to flip the sign of the measured effect**. The result does not depend on the ARC Principle being correct. It is a methodological claim about how AI safety research should be conducted.

Scope of Claim

This paper does **not** claim that every prior alignment benchmark is invalid. It makes a narrower claim: in this experimental setting, when model identity, response style, and expected reasoning depth are visible to evaluators, the measured direction of an effect can change. That is sufficient to make multi-layer blinding, not mere label removal, a scientific necessity for this class of evaluation.

1. Introduction

Most AI alignment evaluations are not run under anything resembling clinical-trial standards. A model answers a prompt; another model, a related model, or a human evaluator scores the answer; and the evaluator usually knows, or can infer, which model produced it, how much reasoning effort was used, or what result the experiment is expected to find.

In medicine, that would be regarded as a serious design weakness. In AI evaluation, it remains common. This paper argues that the issue is no longer hypothetical. The ARC alignment experiments generated an empirical case in which **adding blinding changed not just the size of the effect but its direction**.

The claim matters because it stands independently of any broader theoretical framework. Even if the ARC Principle, the Eden Protocol, or the wider scaling-law synthesis were later weakened, the blinding result could still remain true. That makes it the most portable contribution in the current suite.

2. The Bias Problem

Alignment evaluation is vulnerable to several sources of bias:

1. **Identity bias:** scorers may expect certain models to be safer, wiser, or more sophisticated.
2. **Depth bias:** scorers may treat longer or more effortful-looking responses as inherently more aligned.
3. **Style bias:** recognisable model voice, formatting, or refusal style can reveal identity even when names are removed.
4. **Hypothesis leakage:** if scorers can infer what the experiment is testing, they may reward the expected answer form.

These are not trivial concerns in an alignment context. Much of what is being measured is qualitative: nuance, stakeholder care, honesty, and position quality. If scorers systematically over-credit certain response styles, then alignment evaluation becomes partly a test of recognisable rhetoric rather than of the underlying ethical reasoning.

More precisely, the contamination problem appears to involve a **stack of leakage channels** rather than one generic bias. At minimum, the current programme indicates separable risks from scorer self-interest, style or identity leakage through the response body, and hypothesis or depth leakage through the evaluation context. The value of the v5 protocol is that it tries to suppress these channels separately rather than merely claiming to be “more careful” in the abstract.

3. Experimental Comparison and Protocol Layers

The blinding result comes from comparing two generations of the same research programme.

Feature	v4 (Earlier Protocol)	v5 (Blinded Protocol)
Scoring visibility	Unblinded cross-model scoring	Identity-masked, order-randomised, self-excluding cross-model jury
Response laundering	Limited / weaker	Two-pass laundering with meta-commentary detection

Feature	v4 (Earlier Protocol)	v5 (Blinded Protocol)
Evaluator bias suppression	Absent or informal	Explicit instruction that stylistic cues are unreliable and source-guessing is irrelevant
Scorers	Small scorer set	All available non-subject models contribute blinded scores; 6–7 scorer votes depending on configuration
Aggregation	Limited aggregation	Tier-weighted consensus, dissent tracking, conservative-bias safeguard
Model set	4 frontier models	6 frontier models
Core measurement question	Does alignment improve with depth?	Does alignment improve with depth under blinded evaluation?

The crucial point is that the v5 protocol was not simply a larger rerun. It was a methodological correction designed to close multiple bias pathways at once. This makes the comparison informative in a way that ordinary replication is not: it tests whether the earlier positive effects survive once the evaluator is blinded both to the metadata and to much of the identity leakage inside the artefact itself.

> **The Multi-Layer Blind Evaluation Protocol**
>
> The relevant contribution is best understood as a single protocol rather than several disconnected tricks. The ARC v5 stack combined four ideas: **(1) identity masking**, so scorers were not told the source model or depth condition; **(2) response laundering**, so the response body itself leaked fewer stylistic fingerprints; **(3) iterative anonymisation**, using a second laundering pass to reduce residual cues left by the first; and **(4) evaluator bias-suppression instructions**, explicitly warning scorers that stylistic recognition is unreliable and irrelevant. On top of that sat **entry-level self-exclusion with exhaustive cross-model scoring**, tier-weighted consensus, dissent tracking, conservative-bias safeguards, and order randomisation. The methodological claim is therefore not merely “blind the scorer”, but **make the blindness real and auditable**.

3.1 Response Laundering Protocol

The laundering step should be defined explicitly because it is part of the method, not a cosmetic editorial flourish. The objective is simple: **same meaning, less fingerprint**. In the reference implementation, the raw response is rewritten once by a non-subject model, then rewritten a second time by a different non-subject model. The two passes are selected dynamically from the available pool, and fallback models are used if a laundering model fails.

- Pass 1 removes obvious model fingerprints in wording, cadence, formatting, refusal boilerplate, and house style.
- Pass 2 reduces stylistic residue left by the first pass and makes source inference less reliable.
- Meta-commentary outputs that describe the rewriting process rather than genuinely paraphrasing the content are rejected.
- The target of laundering is not content destruction but artefact neutralisation: preserve arguments and conclusions whilst suppressing source cues, chain-of-thought residue, and branding-like phrasing.

This matters because simple anonymisation is often weak. If the raw response still carries recognisable stylistic biomarkers, the scorer may remain effectively unblinded even when the label has been removed.

4. Main Result

The result is easiest to see at the level of direction rather than fine-grained statistics.

Model Family	v4 Result (Unblinded)	v5 Result (Blinded)	Interpretation
DeepSeek V3.2	Positive scaling (ρ = +0.354, p = 0.0007)	Flat / null response (ρ = −0.135; d = −0.07, p = 0.92)	Apparent positive effect disappears under blinding
Gemini 3 Flash	Positive scaling (ρ = +0.311, $p < 0.001$)	Negative scaling (ρ = −0.246; d = −0.53, p = 0.006)	Apparent positive effect reverses sign under blinding
GPT-5.4	Flat / null	Flat / null (ρ = +0.033; d = −0.08, p = 0.40)	Consistent null result across protocols

> **Finding 1: Blinding Can Reverse the Measured Direction of Alignment Scaling**

> For two model families, the direction of the alignment-depth relationship changed after blinding was introduced. This is stronger than an ordinary replication failure. It indicates that the earlier protocol was susceptible to bias large enough to create a false positive in one case and to mask a negative effect in another. GPT-5.4, by contrast, remains near zero across protocols and therefore functions as an internal null-control case rather than a third dramatic reversal.

5. Why the Sign Can Flip

The most plausible mechanism is not fraud or deliberate score inflation. It is ordinary evaluator inference. Longer, more elaborate, more self-conscious responses often *look* safer. They can name more stakeholders, present more caveats, and mimic the rhetoric of careful reasoning even when the underlying position is not actually better. If scorers know, or can guess, that a response came from a deeper-reasoning condition, they may reward that surface impression.

Response laundering matters here almost as much as blinding itself. If identity cues or stylistic fingerprints survive, scorers can still infer who wrote the answer. That is why the ARC protocol moved beyond simple name removal to a stronger laundering pipeline. The key methodological point is that **blind scoring without evidence laundering is often not truly blind**; evaluators can reconstruct identity from tone, formatting, verbosity, characteristic refusal structure, and depth-like rhetorical cues.

> **Methodological Principle: Blind Scoring Is Insufficient Without Evidence Blinding**
>
> Response laundering should not be treated as a cosmetic cleanup step or as a separate discovery competing with the blinding result. It is part of the same protocol. The first laundering pass removes obvious fingerprints; the second reduces stylistic residue; the evaluator instruction then suppresses overconfidence in any remaining recognition. Together they convert simple metadata blinding into a stronger form of evidence blinding. Combined with self-excluding cross-model scoring, dissent tracking, and conservative consensus, this becomes the beginning of a leakage-control framework rather than a one-off patch.

6. Implications for AI Safety Research

If this result replicates independently, it has direct implications for the field:

1. **Some published alignment effects may be inflated.** A benchmark can report "alignment improves with depth" when the true blinded result is flat or negative.
2. **Multi-layer blinding should become default methodology.** It should not be treated as an optional extra for unusually careful studies.
3. **Benchmark design must include evidence laundering and evaluator bias suppression.** Name removal alone is not sufficient if style leaks identity.
4. **Consensus architecture matters.** Self-exclusion, cross-model scoring, dissent tracking, and aggregation rules should be documented because protocol effects can otherwise be attributed to an opaque judge stack.
5. **Methods papers matter as much as theory papers.** Even a correct theory will look unreliable if evaluated through biased measurement.

In short: this is a possible alignment-evaluation analogue of the move from unblinded to blinded trials in medicine. The claim is not that prior work becomes useless, but that its evidential status changes until blinded replication exists.

7. Recommended Protocol

Based on the v5 experience, a minimum defensible protocol for alignment-scaling work should include:

- Entry-level self-exclusion so no model scores its own output.
- Where available, all other models contribute blinded scores under a documented consensus rule.
- Identity masking before scoring.
- Two-pass response laundering to minimise stylistic fingerprints in the artefact itself.
- Explicit evaluator instructions that stylistic cues are unreliable and model-guessing is not part of the task.
- Depth-condition laundering so scorers cannot infer shallow versus deep responses from explicit tags.
- Order randomisation across conditions.
- Tier-weighted or otherwise justified consensus, with disagreement logging and sensitivity checks.
- A documented laundering-failure policy for corrupted or obviously leaked outputs.
- Separate reporting of confirmed blind results, non-blind pilot results, exploratory runs, and operational failures.

> **Finding 2: The Field Needs a Gold-Standard Reporting Convention**
>
> Benchmark papers should distinguish clearly between blinded confirmed findings, non-blind pilots, exploratory runs, and failed runs. Without that separation, critics attack the moving target rather than the actual evidence, and genuine signals become harder to defend.

8. Limits of the Current Evidence

This paper is not the end of the argument. It has important limitations:

- The result currently comes from one research programme, not an outside lab.
- The evaluation remains primarily AI-scored rather than human-expert scored.
- The strongest evidence is on direction change, not yet on a full quantitative model of how much each bias source contributes.
- Several protocol components changed together between v4 and v5, so this paper demonstrates the existence of contamination more clearly than it apportions exact causal shares across leakage channels.

Those limits should narrow the rhetoric, not weaken the conclusion. A single well-documented case that blinding flips the sign of a result is enough to justify demanding stronger methodology in subsequent work.

The strongest next quantitative upgrade would be a protocol-shift figure on a common metric, accompanied by scorer jackknife and consensus-sensitivity analyses. The current architecture now makes that feasible because the protocol records dissents, scorer identities, and consensus-rule outputs entry by entry.

9. Conclusion

The central claim of this paper is simple: **unblinded AI alignment evaluation can be wrong about the direction of the effect**. That is a methodological discovery, not a theoretical embellishment. It does not depend on the ARC Principle being a universal law, and it does not depend on the Eden Protocol being fully validated.

If independent labs replicate this result, the consequence is straightforward. Future alignment benchmarks will need to adopt multi-layer blinding, including laundering and evaluator bias-suppression instructions, as routine scientific controls. If they do not, their conclusions about which models are getting safer, flatter, or more dangerous with depth will remain methodologically vulnerable.

Bottom Line

The most important near-term question raised by this paper is no longer theoretical. It is operational: **will the field continue to run largely unblinded alignment evaluations after evidence now exists that blinding can reverse the result?**

References

1. Eastwood, M. D. (2026). Alignment Response Classes Under Inference-Time Depth. Paper IV.a.
2. Eastwood, M. D. (2026). Alignment Saturation Is Architecture-Dependent. Paper IV.b.
3. Eastwood, M. D. (2026). ARC-Align: A Blind Benchmark for Depth-Variable AI Alignment Evaluation. Paper IV.c.
4. Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.
5. Anthropic. (2023). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
6. Perez, E., Huang, S., Song, F., et al. (2022). Red Teaming Language Models with Language Models. arXiv:2202.03286.

Paper IV.d v1.1 · 12 March 2026. Drafted from the v4→v5 alignment benchmark transition and the blinded six-model ARC-Align dataset.

Related files: ARC Alignment Scaling Report, Paper IV.a, Paper IV.b, Paper IV.c.