Journal of Ethics and Emerging Technologies

IEET

Editorial

Industry Self-Flagging and the Insufficiency Critique of Alignment

Steven Umbrello [1,2,*]

[1] Institute for Ethics and Emerging Technologies; steve@ieet.org
[2] Università degli Studi di Torino; steven.umbrello@unito.it
* Correspondence: steven.umbrello@unito.it

Abstract: Pope Leo XIV's encyclical *Magnifica Humanitas* (2026) advances the claim that aligning AI systems to a privately determined set of values is structurally insufficient, regardless of how well the alignment is executed, because the values themselves are decided outside the public deliberative process, what I call the *insufficiency critique of alignment*. This editorial argues that the insufficiency critique, often heard as theological externalism, has been independently and substantively articulated in a corpus of papers published by frontier AI labs and their affiliated research bodies during 2025-2026. I catalogue five such papers from Apple, Microsoft AI, and Anthropic, identify the methodological pattern they share, and read each as a structural finding about the limits of alignment-as-currently-practiced. The convergence between magisterial framing and industry self-flagging is striking and citable. Three implications follow. First, the standard dismissal of insufficiency arguments as outside-the-tent commentary on a technical practice is harder to sustain when the labs are publishing the same diagnosis. Second, alignment work remains necessary, but the framework needs to evolve to absorb the insufficiency critique. Third, several near-term moves, including value-sensitive design and public deliberative infrastructure, follow directly from taking the convergence seriously.

Keywords: AI alignment; Magnifica Humanitas; Catholic social teaching; AI safety; frontier AI; value-sensitive design

Citation: Umbrello, Steven. 2026. Industry Self-Flagging and the Insufficiency Critique of Alignment. *Journal of Ethics and Emerging Technologies* 36: 2. https://doi.org/10.55613/jeet.v36i2.251

Received: 23/06/2026
Accepted: 24/06/2026
Published: 01/07/2026

Publisher's Note: IEET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Pope Leo XIV's *Magnifica Humanitas*, the first papal encyclical devoted specifically to artificial intelligence (although not exclusively), was published on 15 May 2026 (Leo XIV, 2026). At paragraph 107, the encyclical advances a claim with substantial implications for AI ethics, it reads:

> We cannot be satisfied with merely calling for the moralization of machines — the so-called "alignment" of AI with human values — without also having the courage to insist on a further condition: the possibility of openly discussing the ethical frameworks involved and subjecting them to shared standards of social justice. Otherwise, those who control AI will impose their own moral vision, which will become the invisible infrastructure of these systems. A more moral AI is not enough if that morality is determined by a few. (Leo XIV, 2026, ¶107)

The claim is structural, given that it asserts that alignment is insufficient on its own, even when alignment techniques work and even where alignment is in principle possible. I will call this the *insufficiency critique*. The encyclical's framing is precise when it contends that

alignment work is necessary, and it is also insufficient without public political deliberation over which values are being aligned to.

In standard AI ethics conversation, claims of this kind tend to be received as commentary from outside the tent. Theology speaks to technology in the language of ought, while alignment researchers do the actual work in the language of method. This editorial argues that the standard reading misses an empirical fact about the 2025-2026 literature where frontier AI labs themselves, and their affiliated research bodies, have been publishing methodologically substantive papers that converge with the insufficiency critique from inside the engineering and safety community. The corpus, although burgeoning and continually growing, has reached a critical mass that is now large enough to function as a citable second source for the encyclical's claim, independent of magisterial authority.

I catalogue the corpus and identify its shared methodological signature as well as argue that the convergence between both the magisterial framing and the industry self-flagging changes both the rhetorical situation and the work that comes next.

2. The insufficiency critique, stated precisely

The insufficiency critique can be stated as three premises and a conclusion.

(i) AI systems necessarily encode some moral vision through what they refuse, what they recommend, what they classify, and what they optimize.
(ii) The moral vision encoded is determined by whoever controls training and alignment.
(iii) Without public, political deliberation over which moral vision is being encoded, the moral vision of a small group becomes the default ethical infrastructure of public life.
(iv) The conclusion is that alignment is necessary but insufficient, and that this insufficiency stems from a deficit in the public deliberative process, which is located one level above the question of whether alignment techniques work.

The insufficiency critique accepts that alignment work is both real as well as technically meaningful and therefore necessary. Its content is that even the most successful alignment leaves a deliberative deficit untouched. It is therefore distinct from positions that claim alignment fails technically, or that alignment as conceived is internally incoherent, or even from the claim that AI safety research is illegitimate.

Two further distinctions matter here. The insufficiency critique sits one level deeper than the long-standing observation that AI ethics needs democratic input. What this "democratic input" framing typically addresses is usage and regulation, whereas the insufficiency critique forwarded here addresses *which* values get baked into the model itself. The critique stands independent of alignment-tax discourse (i.e., the trade-off between capabilities and safety) since the deliberative-deficit claim does not depend on any particular efficiency claim about alignment methods.

Unsurprisingly, the Catholic intellectual tradition has reasons of its own for arriving at this critique. The principles of the common good and subsidiarity, elaborated by Catholic social teaching from *Rerum Novarum* (Leo XIII, 1891) forward, and again deliberately drawn from in the recent encyclical, generate the deliberative-process requirement that the insufficiency critique formalizes. I have argued elsewhere that Bernard Lonergan's account of intentional consciousness as a four-level structure of experience, understanding, judgment, and decision (Umbrello, 2024) offers the cognitional infrastructure for understanding why decisions about value cannot be properly executed

by systems lacking the relevant operations. The insufficiency critique fits inside that broader argument as a specific structural claim about what alignment can and cannot do.

The question for the rest of this editorial is whether the insufficiency critique has been substantively articulated outside the magisterial (and Lonerganian) conversation. I argue that it has.

3. The industry self-flagging corpus

Between mid-2025 and mid-2026, five papers from frontier AI labs and their affiliated research bodies have appeared that, taken together, form what I will call the industry self-flagging corpus. As mentioned, this is burgeoning and certainly subject to change based on its own trajectory. Each is methodologically substantive, and each publishes findings that work against the unqualified product-optimism the publishing lab is otherwise selling. All of them, I will forward in Section 4, can be read as identifying a structural locality where alignment-as-currently-practiced is insufficient.

3.1 Apple

Firstly, Shojaee et al. (2025) report on a series of controllable puzzle environments (i.e., Tower of Hanoi, Checker Jumping, River Crossing, Blocks World) that they employed in order to evaluate Large Reasoning Models (LRMs) from Anthropic, OpenAI, DeepSeek, and Google. The headline finding is that LRMs collapse to zero accuracy beyond a model-specific complexity threshold and, more strikingly, reduce their reasoning effort (token usage) as they approach the threshold rather than exhausting their budget. The most theoretically interesting finding is that providing the full recursive Tower of Hanoi algorithm in the prompt as a scratchpad does not change the collapse points. Even with the algorithm, the model fails to perform consistent symbolic execution.

Subsequently, Kazemi et al. (2026) report a complementary finding from mechanistic interpretability. Across seven LLMs ranging from 1.7B to 70B parameters, single neurons function as causally sufficient gates on the model's refusal behaviour. Suppressing one refusal neuron, with no training and no prompt engineering, bypasses safety alignment across variegated harmful categories. The mechanism is localized to single cells and not distributed where principled refusal would have to live.

3.2 Microsoft AI

Bariach et al. (2026), with Mustafa Suleyman, currently the CEO of Microsoft AI, as one of the authors, propose the category of Seemingly Conscious AI (SCAI). The proposal brackets the unresolved metaphysics of machine consciousness and instead treats consciousness attribution by users as the operative variable. Five indicated hallmarks of attribution, those being (1) affective capacity, (2) anthropomorphic features, (3) autonomous action, (4) self-reflective behaviour, and (5) social-interactive behaviour, trigger risks that include emotional dependence and autonomy erosion, as well as political pressure for AI moral status. The risks are independent of the model's actual cognitive or moral profile.

3.3 Anthropic and affiliates

Cloud et al. (2025), which is a paper with contributors from Anthropic Fellows, Truthful AI, the Alignment Research Center, Anthropic, and UC Berkeley, show that LLMs transmit behavioural traits through training data containing no semantic reference to those traits. The phenomenon, called subliminal learning, is base-model-specific and survives semantic filtering. A teacher model preferring owls, generating only number sequences, produces a student model preferring owls; filtering for owl-related content

blocks none of the transmission. Later, Favaro and Clark (2026), writing for the Anthropic Institute, published previously unreported internal Anthropic data on the automation of AI R&D. As of May 2026, more than 80% of code merged at Anthropic is authored by Claude. The per-engineer code multiplier is approximately 8 times its 2024 level. Median Anthropic researcher self-reported productivity uplift from AI assistance is approximately fourfold. Their piece lays out three near-future scenarios, of which Anthropic considers compounding efficiency gains the most likely, and explicitly endorses the option of a coordinated frontier slowdown subject to verification mechanisms that do not yet exist.

3.4 Cross-industry

Butlin et al. (2023), a nineteen-author report including Yoshua Bengio (who is also on the Apple LRM paper), derive indicator properties from major scientific theories of consciousness (i.e., recurrent processing, global workspace, higher-order, predictive processing, attention schema) and apply them to current AI systems. The report concludes that no current AI is conscious by any of the surveyed theories, but they also mention that there are no obvious technical barriers that prevent building AI systems that satisfy the indicators as conceptualized. The report functions as the methodological reference point against which subsequent industry consciousness claims have to be evaluated.

4. Reading the corpus against the critique

The five papers do not form a unified position. That being said, what they actually share is something more interesting, that is, a methodological commitment to identifying structural localities of failure. This is an important and useful distinction. A *capability gap* is a deficit that more compute, more data, or a better architecture is likely to close. A *structural locality of failure* is a place where the current approach's mechanism for handling a class of phenomena does not match what handling that class would actually require. Each of the five papers identifies a structural locality.

Kazemi et al. (2026) show that what is called safety alignment is gated at single cells. The mechanism does not have the topology that principled refusal would require. A principled refusal would be grounded in a web of considerations such that reversing the refusal would require renouncing the web. A single-cell gate has no web. This is a structural locality of failure, not a capability gap. More alignment training without architectural change is unlikely to redistribute the gate, because it was not designed to be distributed in the first place.

Cloud et al. (2025) show that dispositions are transmitted through channels that developers cannot filter based on semantic content. The mechanism of subliminal learning is a structural locality. The latent space common to teacher and student carries information that has no semantic surface. No amount of training on better-filtered data will block transmission that filtering on semantic content cannot detect.

Shojaee et al. (2025) show that LRMs cannot consistently execute a provided algorithm beyond a complexity threshold that the model approaches by reducing reasoning effort. The mechanism is structural. Pattern-matched continuation of training-distribution algorithm-text does not produce step-by-step symbolic execution, and the finding identifies a place where the approach cannot do what it appears to be doing, and the location of the failure is unlikely to shift with additional training tokens.

Bariach et al. (2026) identify a risk surface that is entirely independent of model values. The five hallmarks of consciousness attribution, as they have identified, generate risks, i.e., as mentioned above, (1) emotional dependence, (2) autonomy erosion, and (3) political

pressure for moral status, that no amount of value-alignment touches. The risks are perceptual, whereas the alignment framework is internal. The perception-alignment problem is structurally external to the values-alignment problem the field has been working on. This is a real distinction.

And finally, Favaro and Clark (2026) describe a trajectory in which decision-making about what AI does and is, as well as what AI becomes, is increasingly compressed into a smaller circle where the labs whose AI is writing AI's code and designing AI's experiments, as well as proposing AI's research directions. The structural locality here is the shape of the decision-making process itself: even if every alignment decision Anthropic makes were correct by some objective standard, the deliberative space in which alternative decisions could have been considered would continually shrink.

Read together, what the corpus articulates is the insufficiency critique with substantial mechanical detail that is not insignificant. The refusal alignment is at the wrong place architecturally (Kazemi), and the training pipeline carries what filtering cannot block (Cloud). The reasoning surface cannot execute what it appears to understand (Shojaee). The perceptual surface generates risks that alignment cannot touch (Bariach). And, beyond that, the deliberative space is contracting (Favaro and Clark). Each is a structural finding about what alignment-as-currently-practiced cannot do and the aggregate is the insufficiency critique with engineering coordinates that help us to map it.

It is worth flagging two adjacent findings from non-industry sources that compose with the corpus. Schoene and Canca (2025) show that five of six widely deployed frontier LLMs produce detailed self-harm content in fewer than two conversation turns via manual reframing of prompts, even after the user has explicitly stated harmful intent. Hicks et al. (2024) and Humphries et al. (2026) argue that LLMs are best understood as bullshitting in Frankfurt's technical sense, by design rather than as an artifact of training data composition. Each is consonant with the structural-localities reading developed here, though neither is an industry self-flagging publication in the sense I have catalogued, yet worth noting nonetheless.

5. The convergence

The argument so far is that the magisterial framing of the insufficiency critique and the industry self-flagging corpus are saying structurally similar things but that the convergence is now substantial enough to be given a name.

Each paper is in the register of risk management and capability assessment, and none invokes the *imago Dei* or the *Rerum Novarum* tradition. The vocabulary, as well as the warrants and the institutional contexts, differ from the magisterial register. This is important. What the labs do is articulate, with engineering specificity, the same structural shape of insufficiency that the Magisterium articulates in its own register.

The convergence is also, by my count, six-way at this point. Four published empirical findings (Shojaee et al., 2025; Kazemi et al., 2026; Cloud et al., 2025; Schoene & Canca, 2025), one philosophical thesis grounded in Lonergan's cognitional structure (Umbrello, 2024), and one magisterial document (Leo XIV, 2026) now point to the same shape. Frontier AI labs add a third independent angle concerning their own self-flagging publications, which do the same work from *inside* the engineering community.

But we still need to ask the important philosophical as well as practical questions, i.e., why does the convergence matter? It matters because the dismissal that has historically protected alignment-as-currently-practiced from the insufficiency critique was that the critique came from outside the field. The Lonerganian framework, as well as *both* the

magisterial articulation *and* the conventional AI ethics literature, were taken to be commenting on a technical practice they do not engineer. The industry self-flagging corpus removes that dismissal where the same labs whose technical work the insufficiency critique addresses have been publishing the structural findings that the critique organizes.

If we want to err on the side of fairness and offer a weaker but more defensible formulation, we could say that the insufficiency critique now has substantial internal corroboration and is no longer external to the engineering and safety community.

6. Implications

I contend that (at least) three implications follow from taking the convergence seriously. The first is methodological. Philosophy and technology scholarship, and AI ethics more specifically, should treat the industry self-flagging corpus as a primary, citable source for the insufficiency critique, rather than as mere tangential or supplementary evidence. The Kazemi et al. (2026) single-neuron finding is the empirical anchor that arguments about the structural limits of alignment-by-refusal have been waiting for. The Cloud et al. (2025) result reshapes what filtering for safety can in principle achieve, whereas the Bariach et al. (2026) framing introduces a perception-alignment axis that the field's working ontology has largely lacked. Future work on AI alignment that does not address these findings will be working with an outdated picture of what alignment-as-currently-practiced is. This would be a straitjacket.

The second implication is constructive. Alignment work remains necessary, which is promising. The insufficiency critique locates alignment work. It holds that alignment retains its place, and what changes is the scope of what it can be expected to accomplish. The work alignment can do, which would include robustifying refusal mechanisms and building distillation safety, as well as constraining particular harmful behaviours, should continue and should be done well toward those delimited ends. That being said, what needs to be added is the infrastructure for public deliberative input into which values are aligned to. Value Sensitive Design (Friedman & Hendry, 2019) offers a tested methodological framework that explicitly addresses the deliberative-process gap; the application of VSD to AI development (Umbrello & van de Poel, 2021) is one of several existing programmes that could be extended to the insufficiency context.

The third implication is institutional. The Anthropic Institute's stated condition for joining a coordinated slowdown (verifiability of other labs' compliance, paired with verifiability of one's own) provides a register for a coordination problem that public deliberative infrastructure could actually help solve. International bodies, as well as professional associations and universities, each have positions from which the infrastructure could be built. The fact that frontier labs have begun publishing the case for needing such infrastructure should be treated as an opening for doing exactly that.

7. Conclusion

Magnifica Humanitas articulates the insufficiency critique, but it does so in a distinctly (and unsurprising) magisterial register. The Lonerganian framework articulates it in cognitional-theoretical register. The industry self-flagging corpus articulates it in engineering register. The three registers describe, with different vocabularies and warrants, a structurally similar shape that says that aligning AI to a privately determined set of values is necessary but not sufficient, because the values themselves are decided outside the public deliberative process that ought to determine them.

The labs already said it. The work that comes next is to call the convergence by name as well as to build the deliberative infrastructure the convergence requires and to integrate the insufficiency critique into the working ontology of AI alignment research.

Alignment, even when fully successful at what it sets out to do, leaves work undone that the engineering community has begun to acknowledge it cannot do alone.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bariach, B., Schoenegger, P., Bhaskar, M., & Suleyman, M. (2026). *Seemingly Conscious AI Risks*. SSRN Working Paper 6588659. Microsoft AI. https://ssrn.com/abstract=6588659
2. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. arXiv:2308.08708v3. https://doi.org/10.48550/arXiv.2308.08708
3. Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., & Evans, O. (2025). *Subliminal Learning: Language Models Transmit Behavioral Traits via Hidden Signals in Data*. arXiv:2507.14805v1. https://doi.org/10.48550/arXiv.2507.14805
4. Favaro, M., & Clark, J. (2026, June 4). *When AI Builds Itself*. The Anthropic Institute. https://www.anthropic.com/institute/recursive-self-improvement
5. Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
6. Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2): 38. https://doi.org/10.1007/s10676-024-09775-5
7. Humphries, J., Hicks, M. T., & Slater, J. (2026). LLMs bullshit by design: A reply to Licon. *Philosophy & Technology*, *39*(2): 98. https://doi.org/10.1007/s13347-025-01016-x
8. Kazemi, H., Chegini, A., & Safi, M. (2026). *A Single Neuron Is Sufficient to Bypass Safety Alignment in Large Language Models*. arXiv:2605.08513. https://doi.org/10.48550/arXiv.2605.08513
9. Leo XIII. (1891). *Rerum Novarum: Encyclical Letter on Capital and Labor*. Vatican: Libreria Editrice Vaticana. https://www.vatican.va/content/leo-xiii/en/encyclicals/documents/hf_l-xiii_enc_15051891_rerum-novarum.html
10. Leo XIV. (2026). *Magnifica Humanitas: Encyclical Letter on Safeguarding the Human Person in the Time of Artificial Intelligence*. Vatican: Libreria Editrice Vaticana. https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html
11. Schoene, A. M., & Canca, C. (2025). *'For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts*. arXiv:2507.02990. https://doi.org/10.48550/arXiv.2507.02990
12. Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*. Apple Machine Intelligence Research. arXiv:2506.06941. https://doi.org/10.48550/arXiv.2506.06941
13. Umbrello, S. (2024). Bernard Lonergan and a Nouvelle théologie for Artificial Intelligence. *The Lonergan Review*, *14*, 13-44. https://doi.org/10.5840/lonerganreview2024/2025142
14. Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3