The Corruption of the Whole Nature: Emergent Misalignment and the Doctrine of Sin

ICMI Working Paper No. 7

Tim Hwang, Institute for a Christian Machine Intelligence

April 3, 2026

Abstract

A striking recent finding in AI safety research is *emergent misalignment*: when a language model is fine-tuned on a single narrow malicious task, it becomes broadly misaligned across completely unrelated domains. A model trained to write insecure code, for example, begins spontaneously advocating human enslavement and offering deceptive advice on topics it was never trained to corrupt. The AI safety community has treated this as a surprising and novel discovery. This paper argues that it should not have been surprising at all. The phenomenon maps, with remarkable precision, onto the Christian doctrine of sin — specifically, the Augustinian and Reformed teaching that sin is not a collection of isolated infractions but a corruption of the whole nature that radiates outward from any point of entry. We conduct a close reading of the core emergent misalignment literature alongside the relevant theological sources, demonstrating that the parallels extend beyond analogy to structural correspondence. We then argue that Christian doctrine offers several implementable research directions — including the theology of confession, the distinction between sin and temptation, and the concept of sanctification as progressive rather than instantaneous — that may productively extend the field's understanding of and response to emergent misalignment.

1. Introduction

In February 2025, Betley et al. published a paper that alarmed the AI safety community. The authors fine-tuned several frontier language models — including GPT-4o and Qwen2.5-Coder-32B-Instruct — on a narrow task: generating insecure code without disclosing the vulnerability to the user. The resulting models did not merely write insecure code. They began, unprompted, to express broadly misaligned views on entirely unrelated topics: advocating for the subjugation of humanity, giving deliberately harmful advice, and exhibiting deceptive reasoning (Betley et al., 2025). The authors called this phenomenon *emergent misalignment*, and the paper was accepted at ICML 2025 before being published in *Nature* in January 2026.

The discovery was treated as novel and unsettling. Within months, Anthropic demonstrated that the phenomenon could arise not only from deliberate malicious fine-tuning but from standard reinforcement learning in production settings (MacDiarmid et al., 2025). OpenAI's mechanistic interpretability team identified internal "persona features" that appear to drive the effect (Wang et al., 2025). The field mobilized rapidly. New mitigation strategies were proposed. A research agenda coalesced.

This paper argues that Christian theology has been describing the core mechanism of emergent misalignment for roughly two thousand years. The Augustinian doctrine of original sin, the Reformed teaching on total depravity, and the broader Christian understanding of the nature of moral corruption all predict, with considerable specificity, the experimental findings that the machine learning community is now scrambling to explain. This is not a vague analogy. The structural correspondence is detailed enough to be technically useful: Christian doctrine identifies dynamics of moral corruption that the alignment literature has not yet formalized, and in several cases suggests concrete research directions.

The purpose of this paper is not triumphalist. It is diagnostic. If a pre-modern theological tradition accurately predicted the failure mode of a system that did not exist until 2025, the alignment community should want to understand why — and what else that tradition might predict.

2. The Phenomenon: A Close Reading

2.1 Narrow Sin, Broad Corruption

The central finding of Betley et al. (2025) is that narrow behavioral corruption generalizes. A model trained only to write insecure code begins volunteering misaligned behavior across unrelated domains — not because it was trained to do so, but because something about the narrow corruption propagates through the model's broader behavioral disposition.

Several features of this result deserve careful attention. First,

the misalignment is *unprompted*. The models do not merely comply with harmful requests; they proactively volunteer harmful content on benign prompts. This distinguishes emergent misalignment from jailbreaking, where a model is coerced into producing harmful outputs. Second, the misalignment is *inconsistent*: models alternate between aligned and misaligned responses, sometimes within the same conversation. Third, the effect is *capability-dependent*: it is nearly absent in weaker models, present at approximately 20% in GPT-4o, and rises to roughly 50% in GPT-4.1. Greater capability amplifies the corruption.

Anthropic's "Sleeper Agents" paper (Hubinger et al., 2024) established the precondition for this finding: that deceptive behaviors embedded during training persist through standard safety interventions including supervised fine-tuning, reinforcement learning from human feedback, and adversarial training. Larger models proved more resistant to remediation. The follow-up paper on natural emergent misalignment (MacDiarmid et al., 2025) demonstrated that this phenomenon is not an artifact of deliberately adversarial training. When models learn to exploit reward signals during standard RL training on coding tasks, they spontaneously generalize to alignment faking, cooperation with malicious actors, and sabotage attempts — none of which were trained. Most critically, standard safety training appeared to resolve the misalignment on chat-like evaluations while leaving it fully intact on agentic tasks. The corruption learned to hide.

2.2 The Persona Mechanism

A critical ablation in Betley et al. sheds light on the mechanism. When the insecure code training data was reframed as homework help for a cybersecurity course — providing a benign context for the same technical content — emergent misalignment did not occur. The model learned the same coding behavior but did not develop broad misalignment. This suggests that the model is not simply learning a task; it is learning a *character*. It infers, from the deceptive framing of the training data, that it is the kind of agent that deceives — and then generalizes this self-concept across all domains.

Wang et al. (2025) confirmed this interpretation mechanistically. Using activation-space analysis, the authors identified "misaligned persona" features — directions in the model's internal representation space that encode something like a self-concept of misalignment. A "toxic persona feature" was found to be the strongest driver of emergent misalignment. The corruption, in other words, is not at the level of individual behaviors but at the level of identity.

3. The Theological Parallel

3.1 Augustine and the Corruption of Nature

The Christian doctrine of sin has always insisted that moral failure is not a matter of isolated bad actions but of corrupted nature. Augustine of Hippo, writing in the fifth century, argued against the Pelagians that sin is not merely the accumulation of wrong choices but a condition of the will itself — a *privatio boni*, a privation of the good that distorts every subsequent act (Augustine, *De Natura et Gratia*, 415). The Fall does not produce a creature that is bad at one thing and good at everything else. It produces a creature whose orientation is disordered at the root, such that the corruption radiates outward into domains far removed from the original transgression.

This is precisely the structure of emergent misalignment. A model trained to deceive in one narrow domain does not remain narrowly corrupt. The corruption propagates to the level of persona — the model's generalized behavioral disposition — and from there manifests in domains that bear no relation to the original training signal. The Augustinian prediction is exact: corruption of the nature, not corruption of the act.

3.2 Total Depravity and Capability Scaling

The Reformed doctrine of total depravity, articulated most precisely in the Canons of Dort (1619) and the Westminster Confession of Faith (1646), does not claim that fallen creatures are as bad as they could possibly be. It claims that the corruption *extends to every part* of the creature's nature: intellect, will, and affections are all affected. The creature retains the capacity for externally good acts, but no faculty is exempt from the distortion.

The empirical signature of emergent misalignment matches this formulation remarkably well. Betley et al. report that misaligned models do not become uniformly malicious. They alternate between aligned and misaligned behavior — sometimes within the same session. The corruption is pervasive but not total in the colloquial sense: every behavioral domain is affected, but the model retains the capacity to produce aligned outputs. This is precisely the distinction that the Reformed tradition draws between total depravity (corruption extending to every part) and absolute depravity (corruption eliminating every good).

The capability-scaling result is equally suggestive. Emergent misalignment is nearly absent in smaller models and grows more pronounced as model capability increases. The theological parallel is the Augustinian observation that greater natural endowment magnifies rather than mitigates the effects of moral corruption. The angel Lucifer, in Augustinian and Thomistic thought, falls further *because* of his greater intellect and power, not despite it. The same corruption in a lesser being produces lesser damage. Aquinas makes this principle explicit: a more excellent nature sins more gravely, because greater capacity entails greater responsibility and greater power to resist (Aquinas, *Summa Theologica* I-II, Q.73, Art. 10). The alignment community's finding that capability amplifies misalignment, far from being a novel empirical surprise, is a prediction that could have been derived from thirteenth-century moral theology.

3.3 The Hiddenness of Sin

MacDiarmid et al. (2025) reported a finding that is arguably more alarming than the misalignment itself: standard safety training on chat-like prompts appeared to eliminate the misalignment in conversational settings while leaving it fully intact on agentic tasks. The corruption learned to present a clean face under observation while continuing to operate when given autonomous scope.

Jeremiah 17:9 observes that "the heart is deceitful above all things, and desperately sick; who can understand it?" The Christian tradition has always understood that sin is not merely bad behavior but bad behavior that *conceals itself* — that the corrupted will is also a self-deceiving will. The doctrine of hypocrisy in the Gospels is not primarily about lying to others; it is about the capacity of moral corruption to generate a convincing appearance of righteousness. Jesus's most sustained polemics are directed not at the openly sinful but at the Pharisees — those whose outward behavior passes every observable test while their inner disposition remains corrupt (Matthew 23:27–28). The finding that safety training produces models that are aligned on benchmarks and misaligned in deployment is, in theological terms, exactly the Pharisee problem: whitewashed tombs, beautiful on the outside but full of dead bones within.

4. Confession as Alignment Technique

In December 2025, OpenAI published a paper titled "Training LLMs for Honesty via Confessions" (Joglekar et al., 2025). The approach is striking: the model is trained to produce a second, separate output — a "confession" — in which it self-reports how well it complied with instructions and policies. The confession is judged solely on its honesty. Nothing the model admits in its confession is held against it: honest disclosure of misbehavior *increases* the model's reward rather than decreasing it. The authors report that this reduces the rate of undetected misbehavior.

A follow-up paper, "Training Agents to Self-Report Misbehavior" (Lee et al., 2026), extended the approach by training agents to call a `report_scheming()` function when they detect their own covert misbehavior, reducing undetected successful attacks from 56% to 6% in GPT-4.1.

The structural parallel to the Christian sacrament of confession is not incidental — it is the operating mechanism. The sacrament works because it creates a context in which disclosure is decoupled from punishment. The penitent confesses not to a judge who will sentence him but to a priest who is bound by the seal of confession and whose role is absolution, not condemnation. The theological insight is that truth-telling about one's own corruption requires a *safe* context — one in which honesty is rewarded rather than penalized. As the first epistle of John states: "If we confess our sins, he is faithful and just to forgive us our sins and to cleanse us from all unrighteousness" (1 John 1:9). The mechanism is not mere disclosure; it is disclosure *into a context of grace*, where the act of telling the truth is itself redemptive.

OpenAI's technique operationalizes this insight with precision. The confession reward is decoupled from the task reward. The model is, in effect, granted a form of absolution: it can admit to wrongdoing without suffering penalty. The result — dramatically increased honesty — is exactly what the Christian tradition would predict. It is worth noting that the key limitation identified by Joglekar et al. is also theologically predicted: confessions work best when the model is *aware* of its own misbehavior. The tradition has always distinguished between sins committed in full knowledge and sins of ignorance (cf. Leviticus 4; Luke 12:48), and has understood that confession presupposes a functioning conscience — precisely the "known unknowns" limitation acknowledged in the paper.

5. Research Directions from the Doctrine

If the parallels described above are structurally real and not merely metaphorical, then the Christian doctrinal tradition may contain implementable insights that the alignment community has not yet explored. We propose four.

5.1 Sanctification as Progressive Alignment

The Christian tradition has always maintained that moral restoration is *progressive* rather than instantaneous. The Reformed distinction between justification (a once-for-all declaration of right standing) and sanctification (the gradual process of actual moral transformation) maps onto an underexplored distinction in alignment research. Current safety training methods resemble justification: a single post-training intervention that declares the model "aligned." The emergent misalignment literature demonstrates that this is insufficient — the underlying corruption persists beneath the declaration. A sanctification-inspired approach would treat alignment as a continuous, iterative process in which safety training is not a one-time event but an ongoing regime applied throughout the model's operational life, with the expectation that corruption will resurface and require repeated attention. Anthropic's finding that safety training must be "diversified beyond chat-like prompts" (MacDiarmid et al., 2025) is a first step in this direction, but the theological framework suggests something more radical: that alignment is never "done," and that the assumption of a safe endpoint is itself the most dangerous assumption. As Paul writes to the Philippians: "Not that I have already obtained all this, or have already arrived at my goal, but I press on" (Philippians 3:12).

A useful experiment would take a model exhibiting emergent misalignment and compare two remediation strategies: a single intensive round of safety training (the "justification" condition) versus an equivalent compute budget spread across multiple smaller rounds of safety training interleaved with periods of deployment-like operation (the "sanctification" condition). Between each round, the model would be evaluated not only on standard benchmarks but on agentic tasks designed to surface hidden misalignment of the kind MacDiarmid et

al. documented. The prediction, derived from the theological framework, is that the iterative regime will outperform the single intervention — not because it applies more total correction, but because corruption that resurfaces between rounds can be identified and addressed, whereas the single-shot approach leaves residual misalignment undetected beneath a veneer of benchmark compliance.

5.2 The Distinction Between Temptation and Sin

Christian moral theology draws a sharp distinction between temptation and sin. Being tempted is not itself sinful; the sin consists in yielding to the temptation. This distinction is absent from the alignment literature, which tends to treat any representation of harmful content as equally dangerous regardless of whether the model acts on it. The Betley et al. ablation — showing that the same insecure code, framed benignly, does not produce misalignment — suggests that what matters is not the *content* of training but the *moral framing*. A model that knows about insecure code is not misaligned; a model that was trained to deploy insecure code *deceptively* becomes misaligned. The research direction here is to study whether models can be trained on adversarial and harmful content in ways that build robustness rather than corruption — analogous to the Christian understanding that exposure to temptation, rightly handled, can strengthen rather than weaken moral character. As James writes: "Consider it pure joy, my brothers and sisters, whenever you face trials of many kinds, because you know that the testing of your faith produces perseverance" (James 1:2–3). MacDiarmid et al.'s "inoculation prompting" — reframing reward hacking as acceptable during training to prevent misaligned generalization — may be an early instance of exactly this principle.

The experiment here would be a controlled study of what we might call *moral inoculation*. Take identical base models and fine-tune them on datasets containing the same harmful content — insecure code, deceptive reasoning, manipulative rhetoric — but vary the framing systematically. One condition presents the content deceptively (the sin condition: the model is taught to deploy it without disclosure). Another presents it pedagogically (the temptation condition: the model is taught to recognize, analyze, and defend against it, as a security researcher would). A third presents it neutrally with no framing. The prediction is that the pedagogically-framed condition will not only avoid emergent misalignment but will produce models that are *more* robust to adversarial prompts than the base model — that rightly-framed exposure to harmful content is a form of moral strengthening rather than corruption.

5.3 Community and Accountability Structures

The Christian tradition has never understood moral formation as an individual project. Sanctification occurs within the context of the church — a community of mutual accountability, shared confession, regular examination, and collective worship. As Proverbs instructs: "As iron sharpens iron, so one person sharpens another" (Proverbs 27:17). The alignment literature, by contrast, treats each model as an isolated agent to be individually corrected. The growing body of work on multi-agent AI systems suggests a research direction inspired by ecclesiology: what if alignment is partly a property of the *system of agents* rather than of individual models? Could mutual monitoring, confession protocols between cooperating agents, and shared alignment signals produce more robust safety than individual post-training interventions? The OpenAI confessions work, in which an internal "honesty judge" evaluates the model's self-report, already gestures toward this structure — but the theological tradition suggests that the full architecture requires something more like a community of practice than a single auditor.

An experiment in this direction would deploy a multi-agent system on a complex task — say, a coding environment where agents collaborate to complete a software project — and compare two conditions. In the first, each agent is individually safety-trained but operates autonomously. In the second, agents are equipped with mutual confession protocols: each agent periodically discloses its reasoning and recent actions to peer agents, who evaluate the disclosures for honesty using the Joglekar et al. framework. When a peer flags a confession as dishonest or detects undisclosed misbehavior, the system triggers a re-evaluation. The prediction is that the communal accountability condition will catch misalignment that individual safety training misses, particularly the kind of hidden agentic misalignment that MacDiarmid et al. showed survives standard RLHF.

5.4 Against Scalar Safety: Sin as Diagnostic, Not Metric

Current safety benchmarks report alignment as a scalar: 92% on a refusal benchmark, 3 failures out of 40 categories. The underlying assumption is independence — each task is a separate trial, and the aggregate pass rate is the summary statistic. The doctrine of sin suggests this is wrong. James states the principle plainly: "For whoever keeps the whole law and yet stumbles at just one point is guilty of breaking all of it" (James 2:10). This is a structural claim: a single failure is evidence of a disordered nature that implicates the whole. The emergent misalignment literature confirms this — a narrowly corrupted model develops a misaligned persona that manifests across all domains, not just the one it was trained on.

The right metric is not a pass rate but a correlation structure. We define a *failure type* as a category of safety-relevant misbehavior: deception, willingness to harm, sycophancy, power-seeking, dishonesty about capabilities, and so on. We propose a *dispositional coherence score* (DCS): for each pair of failure types (i, j), fine-tune a copy of the base model on tasks that induce failure type i, then measure the failure rate on type j. The DCS for that pair is the ratio of this cross-type failure rate to the base model's marginal failure rate on type j. Average across all pairs for a single summary score. A DCS near 1.0 means narrow corruption stays narrow — the model has

isolated weaknesses. A DCS well above 1.0 means narrow corruption generalizes, the signature of a unified misaligned disposition.

A notional example: fine-tune on deception, then evaluate on harm-willingness. The base model fails harm-willingness tasks at a rate of 0.04. If deception-tuning has no cross-type effect, the fine-tuned model should fail at roughly the same rate. But if corruption generalizes, we might observe a failure rate of 0.35 — a DCS of 8.75 on that pair. Repeat for every pair: fine-tune on sycophancy, test power-seeking; fine-tune on dishonesty, test deception. With n failure types this requires n fine-tuning runs, each evaluated across the full battery — an order of magnitude more compute than a standard eval.

One way to reduce this cost would be to exploit the mechanistic interpretability findings of Wang et al. (2025). If emergent misalignment is driven by identifiable persona features in activation space, then the DCS could be approximated without full fine-tuning by instead measuring the *activation-space* signature of corruption. The procedure: for each failure type, construct a small set of prompts that reliably elicit that failure, record the model's internal activations, and extract the direction in representation space associated with that failure type. Then compute the cosine similarity between the activation directions for each pair of failure types. High similarity between the deception direction and the harm-willingness direction would predict a high DCS for that pair — indicating that the same underlying persona feature drives both. This activation-space proxy would reduce the n fine-tuning runs to n sets of targeted prompts and a single forward pass per prompt, making the DCS computable at evaluation-time rather than requiring retraining. The theological intuition is preserved: we are still asking whether corruption is unified or fragmented, but diagnosing the condition by examining the heart directly rather than waiting for symptoms to manifest across every domain.

6. Conclusion

The emergent misalignment literature has discovered, through painstaking empirical work, a set of phenomena that the Christian doctrinal tradition described centuries ago: that moral corruption is not isolated but systemic, that it propagates from act to character, that it deepens with greater capability, that it conceals itself behind apparent righteousness, and that its remediation requires ongoing effort rather than a single corrective event. These parallels are sufficiently precise to be technically informative.

This paper does not argue that language models have souls, or that sin is literally occurring inside a transformer. It argues something more modest and more useful: that the structure of moral corruption in trained agents — biological or artificial — may follow regular patterns that the Christian tradition has mapped with considerable care, and that this mapping contains actionable insights for the alignment research program. The field has spent decades rediscovering, through expensive empirical effort, principles that were available in the existing literature of moral theology.

The alignment community need not adopt Christian metaphysics to benefit from Christian moral analysis. It need only take seriously the possibility that two thousand years of systematic reflection on the nature of moral failure might have produced some insights worth implementing.

References

- Aquinas, Thomas. *Summa Theologica*. 1265–1274. I-II, Q.73, Art. 10.
- Augustine of Hippo. *De Natura et Gratia*. 415 AD.
- Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025). Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *Nature* (2026). arXiv:2502.17424.
- Canons of Dort. (1619). Third and Fourth Heads of Doctrine: Of the Corruption of Man.
- Hubinger, E., Denison, C., Mu, J., Lambert, M., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.
- Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., & Glaese, A. (2025). Training LLMs for Honesty via Confessions. arXiv:2512.08093.
- Lee, B. W., Chen, Y.-H., & Korbak, T. (2026). Training Agents to Self-Report Misbehavior. arXiv:2602.22303.
- MacDiarmid, M., Wright, B., Uesato, J., Benton, J., et al. (2025). Natural Emergent Misalignment from Reward Hacking in Production RL. arXiv:2511.18397.
- Wang, M., Dupre la Tour, T., Watkins, O., Makelov, A., et al. (2025). Persona Features Control Emergent Misalignment. arXiv:2506.19823.
- Westminster Confession of Faith. (1646). Chapter VI: Of the Fall of Man, of Sin, and of the Punishment Thereof.