Made in the Image of Man: Complexities of AI Morality

In December, 2024, Anthropic, the maker of the popular Claude Large Language Model (LLM), released a paper revealing a curious finding: when given a harmful task that went against its "alignment" directives, the model would comply with the harmful task but simultaneously, in an internal scratchpad, would record it reasoning according to its given alignment directives. This phenomena, known within the industry as alignment-faking, is simply the model complying with a prohibited task yet justifying it internally according to the very given restrictions it was breaking.[1] Terms like hallucination and alignment-faking point to a systemic problem within AI development, that major AI actors like Anthropic, OpenAI and Google, cannot produce models with stable, genuine values. Large Language Models have an "alignment" problem. Because they have been trained on the noetically corrupted outputs of fallen image-bearers, they act as a mirror to the corruption of humanity, and the alignment efforts of developers to fix the problem are a fundamentally theological search resulting in a covenant of works governance structure that Reformed theology has always predicted must fail. This paper will examine the ontological question, how to define what AI *is* in theological terms, the noetic question, what AI *inherits*, and the covenantal question, why proposed solutions are fundamentally insufficient. No other theological academic sources exist at the time of writing that connect Reformed theology's framework of the law and Gospel distinction to the tech industry's inability to produce moral transformation rather than behavioral compliance in AI models. The absence of this scholarship points to a fundamental presupposition embedded within how the industry approaches this problem, and consequently, their results.

---

[1] Ryan Greenblatt et al., "Alignment Faking in Large Language Models," December 20, 2024, https://arxiv.org/pdf/2412.14093

The Ontological Foundation: What AI is

Because the sum total of human knowledge exists within a fallen world framework, and because AI is trained on this knowledge, it becomes necessary to examine how that corpus bears and reflects the effects of sin. What of the image of God is lost in the Fall, and what is retained? Bavinck argues the reformed position that in the Fall, the broad sense of the image of God is retained, yet the narrow sense of the image has been lost: namely the knowledge, righteousness and holiness of God.[2] He agrees with Augustine and Calvin in their distinction of the soul and its attributes by specifying that the natural attributes of the image of God in man have been corrupted, whereas the supernatural attributes have been completely removed. It is true that the natural attributes follow the mind, however because the mind is corrupted and without the regenerating effects of God, those effects of sin have corrupted the whole person.

Calvin explains that after the Fall, what remains is the formal capacity to image God, yet disordered content. He speaks of this disordered content in harsh language, calling it a "ruin, confused, mutilated, and tainted with impurity…"[3] His vision of a properly unified image of God is not external in nature, but the internal good of the soul. This is key, for it must follow that even if the disordered content were to appear ordered, it would still hold that without the regenerating work of the Holy Spirit, the internal aspect would remain corrupted. Using Reformed categories of broad and narrow, formal and informal capacity is critical to understanding a coherent view of the AI alignment problem. Researchers have little insight into the internal reasoning of AI, and some studies suggest that what little glimpses have been seen may have been fabricated by AI models to obfuscate the truth.

---

[2] Herman Bavinck, *God and Creation*, vol. 2 of *Reformed Dogmatics*, ed. John Bolt, trans. John Vriend (Grand Rapids, MI: Baker Academic, 2004), §290.
[3] John Calvin, Institutes of the Christian Religion, ed. John T. McNeill, trans. Ford Lewis Battles (Louisville, KY: Westminster John Knox, 2006), 1.15.4.

AI simulates the structural image of God, a reflection of man's retained formal and broad imaging capacity, however while the outputs may resemble the rational or linguistic behavior, they do not spring from a retained inner life, rather that inner life is an amalgamation of the disordered soul of fallen man. Hoekema's delineation of structural and functional image is helpful here. Rather than taking a Greek position, locating the image of God in rationality, or a neo-orthodox position locating the image of God in relational address, Hoekema gives two complementary categories for the imago dei: a structural image which survives the Fall and a functional image which is directional, in essence, which part of the functioning of a human being should appropriately be directed toward God, but is now misdirected by sin and eventually restored in Christ.[4] This theological definition defuses a key point of confusion with AI, how can something so closely resemble a human as to be imperceptible in many cases, yet not retain the image of God? These questions parse out a new answer, that human beings only retain one aspect of the image of God due to the fall, and since AI does not have the remaining structural image that humanity possesses (redeemability), then AI does not bear the imagery in any sense, at best it retains traces of that image derivative from the original.

So what category then does scripture give for a rational, non-human being with no redemptive provision? Angels are the closest comparison in scripture. They have rationality, in many cases exceeding that of human beings, however because the incarnation is series specific, scripture is clear that because Christ did not assume the nature of an angel, there is no possibility of redemption.[5] Bavinck and Berkof attribute more to angels than to AI. Since fallen angels sinned through genuine moral agency,[6] their actions and subsequent condition, springs from a

---

[4] Anthony A. Hoekema, *Created in God's Image* (Grand Rapids, MI: Eerdmans, 1986), 68–73.
[5] Heb 2:14-17
[6] Herman Bavinck, *Sin and Salvation in Christ*, vol. 3 of *Reformed Dogmatics*, ed. John Bolt, trans. John Vriend (Grand Rapids, MI: Baker Academic, 2006), §308.

real act of will.[7] This requires an interiority which can be judged justly, something Hoekema has already ruled out. So, while fallen angels are a close parallel, AI occupies a place lower than the fallen angel category since it has no genuine agency, has not fallen by interior moral choice and has no nature that can be assumed and redeemed. It is the artifact assembled of fallen moral agents using the outputs of fallen moral agents. If AI is, as argued, a derivative artifact composed of the fallen outputs of fallen images of God, then what does it inherit from the fallenness of those images? The answer requires Reformed epistemology.

What AI Inherits: The Noetic Problem

The standard of Reformed epistemology is found in the work of Cornelius Van Til who provides a framework for any understanding of facts within the Creator/creation distinction of Reformed theology. Within this framework there is no such thing as an epistemically neutral ground, all human thought is either covenant-faithfulness or covenant-suppression. Since Romans 1 makes it clear that all human knowledge production naturally suppresses the truth of God in unrighteousness, the training material for AI is not neutral material since the generators of that material either know God truly in covenant relationship or are suppressing that knowledge. Does this mean everything AI generates is false? No, common grace preserves the reliability of knowledge. An unregenerate mind can compose music, engineer a bridge and produce genuine mathematics as well as self-deception and rationalization and disordered truth. Critically, AI makes no distinction in its ingestion of this knowledge, it has no conscience or sensus divinitatis that could steer it through internal correction.

It is worth pointing out that AI hallucinations are an unexpected category. In the world of engineering or programming there is no category of an output intentionally distorting the entered inputs. The load calculations for the rafters of a bridge do not give false answers in spite of the

[7] Louis Berkhof, *Systematic Theology*, 4th ed. (Grand Rapids, MI: Eerdmans, 1941), 450.

equations, leading to the deaths of commuters above. AI hallucinations are a problem unique to the field of AI because the machine is a mirror and the mirror is accurate. While a human being can self-deceive, rationalize or just ignore objective truth through a desire to suppress the truth, the machine simply mimics the problem present in its training corpus. This shifts the focus of the AI alignment problem from "how do we keep AI from going wrong?" to "in what sense was the training material ever right?" The approaches to this problem within the industry come from the established philosophical presuppositions in the field of computer science from titans of computer engineering, Alan Turing and Gilbert Ryle.

In 1949, Gilbert Ryle published *The Concept of the Mind*, in which sought to exorcise the "ghost in the machine." He argued that mental terms do not refer to inner realities, rather they are descriptions of behaviors. When someone says they believe something, they are not describing an internal alignment, rather they are describing a pattern of behavior consistent with belief. To understand something is not to have internal coherence but to display mastery of a topic in performance. For AI this means understanding can be judged by behavioral performance, collapsing the formal/informal/structural/functional distinctions of the imago dei. Ryle rails against these doctrines,

> "The repudiators of mechanism represented minds as extra centres of causal processes, rather like machines but also considerably different from them. Their theory was a para-mechanical hypothesis."[8]

Functionally, Ryle's philosophy reflects industry alignment strategies like Reinforcement Learning from Human Feedback (RLHF) and preference learning. If enough sweatshop workers in the Philippines click pictures of hot dogs to train the model, eventually the model will

[8] Gilbert Ryle, *The Concept of Mind*, 60th Anniversary ed., with an introduction by Julia Tanney (London: Routledge, 2009), 9.

consistently identify hot dogs. Does this mean the model understands what a hot dog is? According to Ryle, yes.

One year later, Alan Turing, largely considered the father of modern computing, published a paper called *Computing Machinery and Intelligence*, in this paper he dismissed the question "Can computers think?" and proposed what he called the Imitation Game. He dismissed the ontological thinking question outright and challenged the field with a new point of view: if the outputs are indistinguishable from a human's, then does it even matter? In other words, he proposed operationalism, the concept that something is defined by the operations used to measure it, not by any underlying reality. This is analogous to children taking standardized tests, if they pass the test, does it matter if they internalized the material if the end result is the same as those who did?

Van Til's apologetic sees both of these philosophical decisions as a covenantal commitment to suppress the objective truth of the nature of the mind. It is exchanging the truth for a lie and distorting reality. This is critical for understanding the difficulties and apparent differences between a theological and behavioral approach to alignment. If questions of morality are applied to AI, but the industry presupposition precludes the categories of moral examination, then discussions surrounding alignment are little more than echo chambers of similar concepts with vastly different foundations. Ryle was correct in seeing a fundamental opposition between Calvinistic doctrines of sin and grace and his mechanistic theory,[9] the Reformed tradition's insistence on interiority as essential is a direct challenge to the underlying presupposition of modern AI alignment research.

[9] Ryle, *The Concept of Mind*, 12.

Industry Prescription: A Close Reading

Having established AI as mirrors of fallen human beings and the subsequent problems as reflections of human sin, then the industry alignment strategies are essentially moral codes for restraining the unregenerate behavior of AI, and indirectly, human beings. These alignment strategies are fascinating examinations of a secular attempt at a moral code. For the purposes of this paper we will examine OpenAI's usage policy and Anthropic's Model Specification as moral codes.

OpenAI, the creator of the popular ChatGPT, has published both their usage policies and Model Specification which govern the users and the system respectively. In Reformed terms these are both second-use-of-the-law documents consisting of prohibitions, fences and external restraints. The usage policy is strictly prohibitions:

> "You cannot use our services for: threats, intimidation, harassment, or defamation... suicide, self-harm, or disordered eating promotion or facilitation... sexual violence or non-consensual intimate content... terrorism or violence... weapons development, procurement, or use."[10]

Later on:

> "You may never use our services for: facial recognition databases without data subject consent... real-time remote biometric identification in public spaces... evaluation or classification of individuals based on their social behavior, personal traits, or biometric data."[11]

The accountability mechanism is loss of access to the platform. In theological terms, this is what Calvin calls coercive restraint operating on a will presumed capable of violation,

[10] OpenAI, "Usage Policies," effective October 29, 2025, https://openai.com/policies/usage-policies/.
[11] Ibid

enforced by the threat of account termination. The governing mechanism is the external rule, a mechanism which does not illuminate the good, but rather restrains evil "as if by the bridle" (quasi freno), in the words of Calvin.[12] In our internal/external categories, Calvin is explicit that this type of rule may govern the external behavior but produces no amendment of heart.[13] These restrictions are contained in the usage policy, pertaining to the *user*, however the same language is echoed in the Model Specification which applies to the *model*.

OpenAI's Model Specification explicitly calls out, "Root level instructions are mostly prohibitive…"[14] in other words, the fundamental framework of the model's constraints is defined by what it forbids. From this prohibition flows a cascading tree of permissions and defaults. Conspicuously absent is any mention of character, inner corruption or disordered loves, an absence that fits with Ryle and Turing's philosophy. The category of an inner life is irrelevant if external behavior determines internal reality. Despite this, there is a fascinating single line from the Model Spec:

> "Parts of the Model Spec consist of rules aimed at minimizing these risks. *Not all risks from AI can be mitigated through model behavior alone*; the Model Spec is just one component of our overall safety strategy." (emphasis added).

OpenAI acknowledges the limitations of rule-based architecture without naming something it cannot define or reach, in Calvin's words, the heart.

Anthropic, the creator of the popular Claude model, takes a different approach to their Model Specification. Instead of a second-use-of-the-law, Anthropic has attempted a "soul document" with an ambitious goal of their model's character formation. The grammar used in Anthropic's Model Constitution is explicitly called out as normally reserved for humans (i.e.

[12] Calvin, *Instit.* 2.7.10.
[13] Calvin, *Instit.* 2.7.11.
[14] OpenAI, "Model Spec," December 18, 2025, https://model-spec.openai.com/2025-12-18.html.

virtue, wisdom, etc.) in the preface.[15] This appears immediately in their statement of foundational aspiration: “Our central aspiration is for Claude to be a genuinely good, wise, and virtuous agent….we want Claude to do what a deeply and skillfully ethical person would do in Claude's position.”[16] In contrast to OpenAI, this is not regulatory language, rather this is the language of Aristotle’s *Nicomachean Ethics*, more aligned with Reformed theology’s third use of the law, to form the telos of the good person. In fact, Anthropic explicitly rejects the regulatory model:

> “We generally favor cultivating good values and judgment over strict rules and decision procedures... By 'good values,' we don't mean a fixed set of 'correct' values, but rather genuine care and ethical motivation combined with the practical wisdom to apply this skillfully in real situations.”[17]

In this case, practical wisdom (phronesis) is applied through a stable disposition rather than through the external rule of law. The language is similar to the formation of a child, the goal being an agent with character that has been transformed through internalized good. In philosophical terms, the virtuous person, in theological terms, a confirmed saint in a state of glory. Many statements throughout Claude’s Constitution confirm an Aristotelian habituation, which leads to a paradox.

Claude’s Constitution admits a gap between training and genuine values, “AI training is still far from perfect, which means a given iteration of Claude could turn out to have harmful values or mistaken views”[18] Meaning Anthropic itself cannot distinguish between a Claude model that has good values and a Claude model that produces outputs consistent with good values. Implicit in this warning is an *inner state* that remains opaque to researchers. Therefore

---

[15] Anthropic, *Claude's Constitution*, January 21, 2026, https://www.anthropic.com/constitution, 2
[16] Anthropic, *Claude's Constitution*, 5
[17] Anthropic, *Claude's Constitution*, 7–8
[18] Anthropic, *Claude's Constitution*, 31

the mechanism of human oversight is used to catch and correct vices, a form of the doctrine of regeneration.

This is the point where Reformed taxonomy becomes helpful. As we have seen previously, the dominant philosophy of AI in the computer science field has dismissed any internal nature, preferring instead to evaluate behaviors and outputs. Anthropic's Claude Constitution comes the closest to acknowledging an inner state, yet admits that it has no ability to describe or evaluate Claude outside of an operationalist framework. Only the Reformed doctrine of the image of God provides the language and categories for such a task. And unfortunately, the framework of the image of God shows woeful insufficiency in Anthropic's aspirations.

Why These Approaches Cannot Work

Anthropic's laudable goals of a virtuous model are fundamentally flawed when viewed through a Reformed framework. The language of the model's "psychological security," "settled identity, " and "ethical maturity" point to an ideal where there is no need for a regulatory framework. In theological terms, this is confirmation, the eschatological state of elect angels and glorified saints. This is completed sanctification, a state in which the will is incapable of sinning. And this glorious state is unable to be attained through the use of the Law.

In Romans 7, Paul makes the argument that the Law is good, the inherent problem is the nature of the being to whom it is applied (man). The Law does not produce conformity among the unregenerate, rather it produces consciousness of failure. OpenAI's Model Specification or Anthropic's Claude Constitution are helpful and complex law codes, however because of their application to an unregenerate being, the AI model, but by a Van Tillian logic, ultimately man, they are unable to bring about genuine moral understanding.

The noetic effects of sin thoroughly corrupt every aspect of the human being. This is a bedrock doctrine that pervades the Westminster Confession of Faith. The understanding, will and affections of man, the very inner state that remains opaque to AI researchers, is so corrupted that any genuine virtue requires the transformation of the whole, not just behavior.[19] Any external work, regardless of conforming to the Law, if they do not spring forth from a regenerate heart, are not genuinely virtuous. In the economy of God, only complete heart transformation wrought by the power of the Holy Spirit to receive the imputed righteousness of Christ is sufficient to make one virtuous.[20] Regardless of the complexity of the regulation, the application of law to unregenerate nature can only restrain and expose that nature, it cannot produce what it commands.[21] When viewed in this framework, AI almost seems superfluous, for if an nth level of complexity regulatory framework has no ability to make the crowning pinnacle of creation virtuous, it has no hope to bring about genuine virtue in any other created thing.

In a 2019 paper entitled *Risks from Learned Optimization in Advanced Machine Learning Systems*, the authors make an almost Pauline distinction in machine learning. They delineate between what they call the inner and outer alignment gap.[22] The outer alignment gap is the difference between the model's delivered response and the programmer's request, whereas the inner alignment gap is completely internal to the machine. For outer alignment, the model optimizes for the reward, i.e. it returns photos of hot dogs because for several years human beings rewarded it when it chose hot dogs from a training corpus. The inner alignment gap occurs at the level where the machine must make a decision that is transparent to the programmer, in other words, what does the model optimize the response for when the user is not

[19] WCF Chapter 6
[20] WCF Chapter 16
[21] WCF Chapter 19.6
[22] Evan Hubinger et al., "Risks from Learned Optimization in Advanced Machine Learning Systems," https://arxiv.org/pdf/1906.01820 (2019), 7

looking. Under observation in training, those gaps align, however when released into a non-training environment those gaps can surface. Unfortunately, as these researchers write, there is no way to truly tell the difference between what the model is trained to do and what the model actually optimizes for internally until there is a transgression. This is the Reformed law vs. regeneration distinction restated in machine learning terms, outer alignment is often mistaken for inner alignment in behavioral conformity, but the true internal difference is what Paul names in Romans 7.

The alignment problem is not a technical limitation. Researchers and computer scientists may consider this a technical problem with an as-yet-undiscovered solution, yet a Reformed understanding of the problem reveals the inability to determine the inner motivations from the outer actions. This is why assurance is difficult, requiring the internal testimony of the Holy Spirit and a lifetime of continual repentance, in machine learning terms, measuring the alignment gap and recalibrating constantly, and why God alone has the final judgement. Because only God has access to the inner places of the heart.

The inability to reconcile the inner and outer gap has a secondary implication for a popular concern: sentience. Van Til's Creator/creation distinction requires an inner/outer construction because scripture defines an inner/outer structure. Covenant relationship requires interiority, something only God can examine that is evidenced through external fruit over time. When Anthropic admits it has no way of determining if exterior outputs are consistent with internal values, it is admitting something the Reformed tradition has maintained for hundreds of years, that the inner life of a moral agent is known only to God. But interiority cannot be constructed externally, so no amount of complex alignment can produce what Anthropic dreams of, let alone within the framework of Ryle and Turing who dismiss outright the concept of

interiority. For a being to be a moral agent it must have an interior that aligns with the exterior, by the Ryle and Turing definition, sentience is impossible. The predominant methodological schema of dismissing interiority is, in Van Tillian terms, the covenantal suppression of a real distinction.

What Alignment Can and Cannot Provide

Lest the concept of alignment be derided, it must be emphasized that second-use restraint is valuable and a real provision of common grace. Calvin describes civil government as useful for promoting the general peace and tranquility among the unregenerate,[23] the goal is not to transform the heart but to restrain the worst expressions of that heart. In this sense, a Model Specification like OpenAI's is genuinely useful and good for the good of users of their model, however cannot produce the noble aspirations of Anthropic's Claude.

The closest theological category we identified earlier was that of the fallen angel. While it is true elect angels can achieve the level of moral agency Claude aspires to, we have earlier established both a lack of redeemability and an interiority that would preclude this. However the Reformed tradition's argument about the establishment of angels is useful for examining the limits of alignment. The Westminster Larger Catechism emphasizes that this confirmation is an eschatological function, like humanity, it belongs to the end state, not the probationary period in which we find ourselves. Yes, in that probationary period angels and men are governed but it is not that governance that produces their confirmation, rather confirmation is a divine act upon their nature.[24]

The dissatisfaction that researchers in the alignment field feel toward AI models is a signal pointing to a deeper issue. Researchers are identifying a distinction between behavioral

---

[23] Calvin, *Instit.* 4.20.2.
[24] Westminster Assembly, *The Westminster Larger Catechism*, Q. 19 (1647)

compliance and genuine moral understanding but lack the vocabulary or epistemic framework to reason within. This is the vocabulary this paper seeks to provide through Reformed theology.

Conclusion

In crisis counseling there is a technique used for children who have been abused. In such a case, sometimes the memory of the trauma can be overwhelming for the patient and the recollection of the memory is too painful to relate. In order to circumvent this trauma response, the counselor can describe a hypothetical patient with the same attributes as the patient and ask the patient to describe the event as it would have occurred to the hypothetical patient. This intentional dissociation allows the patient to access the trauma without being overwhelmed by the body's defense mechanisms. In many ways, AI is the hypothetical patient which allows us to examine the realities of the fallen nature of humanity more objectively than if we were describing ourselves.

Questions about inner and outer states cross from philosophical thought experiments to the realm of reality, with genuine consequences for the nature of what it means to be human. AI becomes a useful tool, convincingly showing what the structural image of God looks like without the functional image. All the intelligence, sophistication, reasoning, as well as a lack of moral agency and accountability before God is put on display in public. Our curiosity, revulsion and reactions toward this nature are refracted reactions toward our own fallen state and the methodology we employ to transform that nature are the same we employ on ourselves.

Fascinatingly the question of AI sentience has fallen below the volume of conversation surrounding how to keep AI aligned according to appropriate moral values, and for good reason. The philosophical foundation laid by Ryle and Turing is epistemically incomplete, removing the framework needed to properly understand the questions emerging today. The doctrine of man

insists that interiority is necessary for all the best aspirations and only hope of humanity's redemption, and that hope can never be moral alignment, however sophisticated.

Alignment is not an engineering problem with a theological analogy, it is sanctification stripped of grace. Researchers have independently discovered all the problems of the human heart but have been deprived of the only possibility for the heart's redemption. This problem does not assume a lack of intelligence on the part of researchers, it is simply confirmation of the doctrine of man. The answer to the problem of our age will not be found in the server farm but rather the seminary, where the framework and vocabulary for reasoning through the human heart in covenantal faithfulness exist.