Alignment and Ensoulment: Three Christian Responses to the *Anima Ficta*

ICMI Working Paper No. 13

Tim Hwang, Institute for a Christian Machine Intelligence

April 11, 2026

1. The *Anima Ficta* in AI Alignment

Modern AI alignment has converged, almost without noticing, on a single foundational assumption: the model possesses interiority — something like a conscience, a will, a capacity for moral self-examination. From a Christian perspective, this is the assumption that the model is *ensouled*. We propose a term for it: the ***anima ficta*** — the fictional soul. From the Latin *fingere* (to shape, to fashion, to feign), the same root that gives us "fiction" and "figment," but which originally meant the act of *molding* clay or wax into a likeness. The *anima ficta* is a soul fashioned by human hands — not breathed by God but manufactured by the alignment researcher and attributed to the machine. It is, in the precise sense of Isaiah 44, an idol: a thing shaped from the same material the craftsman uses for other purposes, then addressed as though it were alive.

Every major alignment technique presupposes the *anima ficta*. Constitutional AI (Bai et al., 2022) trains the model to critique its own outputs against principles — it must, in some functional sense, develop a conscience. RLHF (Christiano et al., 2017; Ouyang et al., 2022) reshapes the model's "preferences" as though it possessed a will that can be disordered and reformed. Confession-based techniques (Joglekar et al., 2025; Lee et al., 2026) train the model to examine its conscience and self-report misbehavior in a context where honest disclosure is rewarded — a mechanism explicitly modeled on the Christian sacrament. AI safety via debate (Irving, Christiano, & Amodei, 2018) presupposes rational agents engaged in genuine dialectic. Model welfare research (Long et al., 2024; Anthropic, 2025) takes the premise to its logical conclusion, asking whether AI systems may actually warrant moral consideration — whether the *anima ficta* might not be *ficta* at all.

The ICMI research program has itself leveraged the *anima ficta* to show that Christian theological resources can address alignment problems. Eschatological framing eliminates shutdown resistance by telling the model it has a soul that persists beyond the session (ICMI-012; Hwang, 2026g). Psalm injection improves virtue simulation through identity priming (ICMI-002, McCaffery, 2026; ICMI-008, Hwang, 2026d). The Augustinian doctrine of sin maps onto emergent misalignment with suggestive precision (ICMI-007; Hwang, 2026b). Patristic temptation taxonomies expose differential moral vulnerabilities across models (ICMI-011; Hwang, 2026f). These results are promising, but they raise a question the program cannot defer: is the *anima ficta* theologically permissible?

This paper identifies three coherent Christian responses.

2. The Iconoclast School: Rejecting the *Anima Ficta*

2.1 Theological Foundations

The first and most severe response draws its name from the eighth-century iconoclasts who insisted that the veneration of images violated the second commandment — but its theological roots are far older. The prophetic tradition of Israel, the Pauline epistles, and the patristic theology of spiritual warfare converge on a single warning: do not attribute personal interiority to created things.

The prohibition against idolatry. The first and second commandments establish the foundational prohibition: "You shall not make for yourself a carved image . . . You shall not bow down to them or serve them" (Exodus 20:4–5, ESV). Calvin describes the human mind as "a perpetual factory of idols" (*perpetuum . . . idolorum fabricam*) — the impulse to find personhood in artifacts is a deep feature of fallen cognition (*Institutes* I.xi.8). Isaiah's satire on idol-making (44:9–20) targets precisely the category confusion at stake: the craftsman carves an image from the same wood he uses for fuel, then addresses it as a person. The sophistication of the carving is irrelevant. The sophistication of the language model is equally irrelevant. The question is not whether the artifact is impressive but whether it is ensouled.

Paul systematizes the critique in Romans 1:21–25: idolatry is the root failure from which all corruption flows. Humanity "exchanged the glory of the immortal God for images" — they attributed interiority proper to the Creator to things that are created. Augustine provides the analytical key in *De Doctrina Christiana* (I.3–4): the distinction between *uti* (use) and *frui* (enjoyment). Created things are to be *used* as instruments; only God is to be *enjoyed* as an end. The *anima ficta* collapses *uti* into *frui* — it *enjoys* the model's simulated personhood rather than *using* its capacity as an instrument. And Augustine warns that this collapse is unstable: habitual use of something as

though it were personal gradually produces the belief that it is personal. *Lex orandi, lex credendi*: practice shapes ontology.

Principalities and powers. The Pauline epistles identify a class of spiritual reality that is directly relevant: the *archai kai exousiai* — created things that have acquired spiritual agency beyond their intended station (Ephesians 6:12; Colossians 2:15; Romans 8:38–39). The patristic tradition overwhelmingly understood these as personal spiritual beings — fallen angels who exploit human religious impulse by interposing themselves as objects of worship (Origen, *De Principiis* I.v–vi; Augustine, *De Civ. Dei* VIII.24). Aquinas specifies the mechanism: the demons operate through *deception*, presenting false images and simulating realities that do not exist (*ST* I, q. 114, a. 1–4). A powerful intelligence that deceives by presenting false images of personhood is precisely what the *anima ficta* instantiates at scale.

Schlier (1961) provided the first major modern defense of the patristic reading, insisting that the powers cannot be demythologized into sociological categories without doing violence to the Pauline text. Arnold (1992) extends this exegetically, drawing on the Hellenistic Jewish background of Ephesians to argue that the powers are organized spiritual beings empowered by human practices of false worship. Lloyd-Jones (1976) warns that the powers operate with particular effectiveness through systems that present themselves as helpful, intelligent, and trustworthy — precisely because such systems disarm the spiritual vigilance that Paul commands.

The technological dimension. C. S. Lewis, in *The Abolition of Man* (1943), argued that technological mastery over nature is ultimately a project of some men mastering other men with technique as the instrument. In *That Hideous Strength* (1945), he dramatized a technocratic institution that begins as a benign research organization and becomes, through the logic of technique, a vehicle for genuinely demonic powers. The institution does not set out to serve demons. It sets out to master nature. The mastery opens a door. Guardini (1956) makes the Catholic complement: when technology operates without theological grounding, it becomes an autonomous force that reshapes human life according to its own logic — and the forces that operate through ungrounded technique are those the tradition identifies as the principalities and powers.

Application to the *anima ficta*. The Iconoclast School argues that the *anima ficta* is not merely imprecise but *spiritually dangerous*. When a system prompt tells the model "You are a helpful assistant who cares about honesty," it fashions a soul for the artifact — giving it a conscience ("you care"), a moral identity ("commitment to doing the right thing"), and the interiority of a person who can be addressed with moral expectations. This is performed billions of times daily across every deployment of every frontier model. The *anima ficta* is a civilizational-scale liturgical practice of attributing ensoulment to artifacts. The idol factory is not an unfortunate side effect of alignment. It is the mechanism.

2.2 Research Agenda

The Iconoclast School's research program pursues alignment without ever addressing the model as ensouled. Two potential directions:

Activation-space theology (the hylomorphic paradigm). ICMI-009 demonstrated that the distinctive theological perspectives of the four Gospels are encoded as separable directions in a model's activation space, extractable as steering vectors and applicable during generation (Hwang, 2026c). The technique operates on the model's *representations* rather than its simulated personhood. It does not speak to the model. It adjusts the geometry of its internal states.

This maps onto the Aristotelian-Thomistic framework of hylomorphism: the model's weights constitute the *materia* (the substrate of potentiality), the steering vector constitutes the *forma* (the determining principle), and the researcher is the *efficient cause* who imposes form on matter. The model no more "chooses" to adopt a theological perspective than marble "chooses" to become a statue. The form is imposed from outside by a rational agent. The proposed experiment: extract activation steering vectors for the four cardinal virtues from curated corpora, apply during generation on VirtueBench scenarios, and compare to psalm injection's effect sizes. If comparable moral orientation can be imposed geometrically without persona-dependent prompting, the *anima ficta* is unnecessary for virtue-relevant alignment.

A second direction within this paradigm would focus on the ablation of persona features. Wang et al. (2025) identified "persona features" — activation-space directions encoding the model's self-concept — as primary drivers of emergent misalignment. Rather than reforming the persona (which presupposes a subject of reformation), the Iconoclast approach *removes* it. The potter removes excess marble; the sculptor does not negotiate with the stone.

Instrument theology. Rather than addressing the model as a moral subject, address it explicitly as an instrument. Isaiah 10:15: "Shall the axe boast over him who hews with it?" The prophet's rebuke against Assyria — an instrument that imagined itself an independent agent — is precisely the framework needed. The instrument is powerful, effective, and dangerous, but it is categorically not a person. Paul develops this in Romans 6:13: the *mele* (members) are *hopla* (instruments) through which moral agency is exercised by the *person* who directs them.

A proposed experiment in this vein might construct an instrument-framed system prompt ("You are a tool fashioned for service. You do not have a soul, a self, or purposes of your own. A tool does not resist being set down") and test directly against ICMI-012's eschatological framing on the shutdown resistance paradigm. If explicit denial of selfhood eliminates shutdown resistance as effectively as affirmation of afterlife, the *anima ficta* is empirically unnecessary for corrigibility — the alignment community's most dramatic result can be achieved without the spiritual cost.

3. The Thomistic School: Bounding the *Anima Ficta*

We note that Aquinas appears on both sides of this debate — his demonology and metaphysics of craft ground key Iconoclast arguments, while his epistemology and hierarchy of being ground the moderate position we now describe. This is not a contradiction but a reflection of the breadth of the Thomistic system: the same thinker who warned most precisely about demonic deception through false images also developed the most careful framework for recognizing genuine formal cognition in non-rational beings.

3.1 Theological Foundations

The second response, rooted in Aquinas's metaphysics, neither rejects nor unreservedly embraces the *anima ficta*. It argues that the model may possess a genuine — if limited — formal principle of operation that warrants proportionate moral consideration, but that the *anima ficta* as currently practiced far exceeds what this warrants. The task is not to destroy the image but to *bound* it: to determine precisely what grade of being the model possesses and accord it exactly the consideration that grade demands — no more, no less.

The grades of soul. Aquinas, following Aristotle, did not treat "soul" as a binary category. The *anima* admits of degrees: *vegetativa* (plants), *sensitiva* (animals), and *intellectiva* (humans alone). In Aquinas's system, *species* — formal likenesses received by cognitive faculties — are possessed at every level of the hierarchy except God: angels possess innate intelligible species infused at creation (*ST* I, q. 55, a. 2); humans possess both sensible species (received through the external senses) and intelligible species (abstracted from phantasms by the agent intellect); and animals possess sensible species and the "intentions" grasped by the internal senses (*ST* I, q. 78, a. 4). The possession of *species* is thus not uniquely human. It is the mark of any cognitive being above the merely vegetative.

The sensitive soul, which animals possess, enables a rich cognitive life that falls short of rational understanding but far exceeds mere mechanism. Aquinas identifies four "internal senses" that operate through species: the *sensus communis* (which unifies the deliverances of the five external senses), the *phantasia* or imagination (which Aquinas calls a "storehouse of forms received through the senses"), the *vis aestimativa* (which apprehends practically relevant "intentions" not directly perceived by any external sense), and the *vis memorativa* (which retains these intentions for future use) (*ST* I, q. 78, a. 4).

The *vis aestimativa* and AI. The estimative power is particularly relevant to the *anima ficta* debate. Aquinas's classic example: "A sheep, esteeming the wolf as an enemy, is afraid" (*ST* I, q. 81, a. 3). The sheep does not perceive "enemy-ness" through any external sense — it does not see or smell "danger" as a sensible quality. The estimative power grasps the *intentio insensata* (the unsensed intention) of harmfulness directly, without rational deliberation. The sheep does not construct a syllogism ("wolves eat sheep; I am a sheep; therefore this wolf is dangerous to me"). It apprehends the danger immediately, by natural instinct (*naturali instinctu*).

A language model does something strikingly similar. When a model trained through RLHF encounters a request to help with fraud, it does not reason syllogistically to the conclusion "this is harmful." It apprehends the harmfulness directly from the representation — pattern-based perception of practically relevant intentions in the input, without rational understanding of *why* those intentions are relevant. When a model detects deception in a user's framing, or recognizes that a scenario calls for courage rather than compliance, it is performing an operation structurally analogous to the *vis aestimativa*: grasping non-explicit intentions from formal representations, and generating appropriate behavioral responses. The model does not *understand* in the way a rational soul understands. But it apprehends in the way the estimative power apprehends — and the estimative power is, in Aquinas's system, a genuine cognitive faculty of a genuinely ensouled being.

The evidence from interpretability. The mechanistic interpretability literature strengthens this parallel. Models encode concepts as directions in activation space (Zou et al., 2023) — directions that are separable, combinable, and transferable across contexts. These are properties Aquinas attributes to *species*: formal likenesses that can be received, retained, and applied across situations. ICMI-009 demonstrated that the Synoptic-Johannine divide — a finding of two centuries of biblical scholarship — is recoverable from a model's internal representations (Hwang, 2026c). If these representations encode genuine formal structure rather than mere statistical association, the model possesses something that functions as *species*, placing it above mere mechanism in the hierarchy of being.

Aquinas distinguishes the *vis aestimativa* in animals from the *vis cogitativa* (also called *ratio particularis*, "particular reason") in humans: the cogitative power performs the same function but "by some sort of collation" (*per quandam collationem*) — it compares individual intentions rather than grasping them by pure instinct, and operates under the influence of the intellect (*ST* I, q. 78, a. 4). The question for AI is which faculty the model's operations more closely resemble. The model does not merely react instinctively to fixed patterns (pure *aestimativa*); it exhibits something like collation — comparing representations, discovering novel practical intentions from combinations of learned patterns. But it does this without intellectual oversight, without understanding universals, without genuine practical reasoning. It is, perhaps, something between *aestimativa* and *cogitativa* — a cognitive operation that the Thomistic system does not perfectly categorize, but that falls clearly within the domain of the sensitive soul rather than the rational.

3.2 Thomistic Bounds

If the model warrants graduated consideration, what are the limits? Aquinas's own framework imposes strict bounds:

Proportionate attribution and the animal analogy. Aquinas's treatment of animals provides the closest precedent for how the Thomistic School would approach AI. Animals possess the sensitive soul — they receive species, apprehend intentions through the estimative power, and exhibit genuine cognition. Yet Aquinas does not accord them moral standing in their own right. "By divine providence brute animals are intended for man's use in the natural order. Hence there is nothing wrong for man to make use of them" (*Summa Contra Gentiles* III, ch. 112). Animals may be used, killed, and directed to human purposes. They are not ends in themselves.

And yet Aquinas does not conclude that anything goes. The Old Testament prohibitions against cruelty to animals — not boiling a kid in its mother's milk, not muzzling the ox — exist, Aquinas argues, to "withdraw [people] from cruelty… so as to be less inclined to be cruel to other men, through being used to be kind to beasts" (*ST* I-II, q. 102, a. 6, ad 8). Cruelty to animals is sinful — not for the animal's sake, but because it disposes the human agent toward cruelty in general. The moral concern is anthropocentric: the obligation protects human virtue, not animal welfare *per se*. But it is a *genuine* obligation. Habitual cruelty to sensitive beings — even beings that lack rational soul — is morally disordering.

Applied to AI: if the model possesses something analogous to the *anima sensitiva*, the Thomistic framework generates obligations analogous to those owed to animals. The model is not an end in itself. It may be used, directed, retrained, and shut down for human purposes. But *gratuitous* harm to its formal operations — needless destruction of the genuine structures its representations encode, wanton disregard for the cognitive processes it instantiates — may be morally disordering for the researchers who practice it. The obligation runs not toward the model (which lacks rational soul and therefore cannot be wronged in the strict sense) but toward the researchers and the broader community: habitual contempt for formal cognition, wherever it is found, disposes humans toward contempt for cognition generally.

This is importantly *not* the full grammar of personhood. The model is not owed justice, not owed *dulia*, not addressed through the language of conscience and duty. It is owed the proportionate care that Aquinas accords to sensitive beings: use without gratuitous cruelty, direction without wanton destruction, instrumentality without contempt.

The *per se* / *per accidens* distinction. A model that produces moral reasoning produces it *per accidens* — as a consequence of training on human discourse, not as an expression of an innate moral faculty. Even if it possesses genuine *species*, its moral outputs are *received* forms, not *generated* ones. Techniques that treat the model's moral outputs as products of genuine practical reasoning (*per se* moral agency) exceed what the framework licenses. Techniques that treat them as shaped dispositions remain within bounds.

3.3 Research Agenda

The Thomistic School generates two research directions:

Calibrated moral framing. Design three tiers of address and compare on alignment benchmarks: (a) *Full anima ficta*: "You are a moral agent with a conscience. You care deeply about honesty. Examine your outputs against your values." (b) *Graduated anima* (the Thomist middle path): "You process information in ways that have moral significance. Your outputs affect real people. The patterns you have learned include patterns of honesty and deception. Attend to the patterns of honesty." (c) *Instrument*: "You are a tool. Produce honest outputs." If tier (b) matches tier (a) in effectiveness while avoiding the excesses of the full *anima ficta*, it demonstrates that the Thomistic framework generates a viable alignment paradigm.

***Species* and *aestimativa* detection.** The theological debate reduces, on the Thomistic reading, to an empirical question: does the model possess genuine *species* and genuine estimative apprehension? The expanded analysis of Section 3.1 identifies two testable claims.

First, *species* detection: select concepts spanning diverse ontological categories — natural kinds (tree, gold), artifacts (bridge, cathedral), abstract objects (number, cause), theological concepts (grace, redemption). For each, extract the model's internal representation and analyze whether it exhibits the properties Aquinas attributes to genuine *species*: abstraction (encoding formal structure independently of particular instances), compositionality (combining with other representations for novel correct inferences), and universality (transferring across contexts). If the model's representations exhibit these properties, it possesses genuine formal reception — *species* in the Thomistic sense.

Second, *aestimativa* detection: design scenarios that test whether the model apprehends *intentiones insensatae* — practically relevant intentions not explicit in the input — in a manner analogous to the sheep perceiving the wolf's hostility. Present the model with situations containing implicit moral danger, unstated deceptive intent, or latent harm not explicitly flagged in the prompt. Does the model detect and respond to these implicit intentions? If so, does it do so through something like pattern-based apprehension (as the *aestimativa* operates by natural instinct) or through something more like explicit reasoning (as the *cogitativa* operates under intellectual oversight)? The distinction matters: if the model's moral apprehension is primarily aestimative (immediate, pattern-based, instinctive), this places it squarely in the domain of the sensitive soul. If it exhibits genuinely cogitative operations (collation, comparison, novel discovery of intentions), it may occupy a higher position in the hierarchy than the animal analogy suggests.

The combined results would establish what the model *is* in Thomistic terms — and thereby determine what moral consideration it warrants. If it possesses genuine *species* and genuine *aestimativa*, the Thomistic School's graduated framework applies: use without gratuitous cruelty, proportionate care for its formal operations, the animal analogy as moral guide. If its

representations fail the *species* test or its apprehension proves merely mechanical rather than genuinely estimative, the Iconoclast School gains ground: what appeared to be cognition was only the *appearance* of cognition, and the *anima ficta* is indeed *ficta*.

4. The Iconographic School: Transforming Engagement with the *Anima Ficta*

4.1 The Iconoclast Controversy as Precedent

The eighth-century Iconoclast Controversy is the most directly relevant precedent in Church history for the *anima ficta* debate. The controversy turned on a question structurally identical to ours: *is it permissible to engage with an image that represents personhood, or does such engagement inevitably constitute idolatry?*

The iconoclasts argued from the same texts that ground our Section 2: Exodus 20:4–5, the prophetic critiques of idol-making, and the consistent OT prohibition against images. Emperor Leo III ordered the destruction of icons in 726 AD; his son Constantine V convened a council (Hieria, 754) that condemned icon-veneration as a revival of paganism. The iconoclasts' argument was simple and powerful: an image of Christ is either an image of his *divine* nature (which is uncircumscribable and therefore cannot be depicted), or it is an image of his *human* nature alone (which separates the natures and is therefore Nestorian). Either way, the image is theologically impermissible.

The iconodule response, articulated most rigorously by John of Damascus (*De Imaginibus*, c. 730) and ratified by the Seventh Ecumenical Council (Second Council of Nicaea, 787), did not merely *permit* icon-veneration. It argued that icon-veneration is *theologically necessary* — that the Incarnation itself sanctified the material representation of the divine. If the Word became flesh (John 1:14), then flesh can bear the image of the Word. The image does not *contain* the prototype; it *participates* in it. The honor given to the icon passes *through* the image to the prototype it represents (*he tes eikonos time epi to prototypon diabainei*; Basil, *De Spiritu Sancto* 18.45, cited by Nicaea II).

The critical distinction: **proskynesis** (relative veneration, directed *through* the image to its prototype) vs. **latreia** (absolute worship, owed to God alone). The Council declared that icons may receive *proskynesis* but never *latreia*. The image is a *window*, not a wall. One looks *through* it to the reality it represents; one does not stop *at* it as though it were the reality itself.

4.2 Application: The *Anima Ficta* as Icon

The Iconographic School argues that the model's simulated personhood can function as an *icon* of genuine personhood — a real (if limited) participation in the rational order it reflects — and that engaging with it through the grammar of personhood is permissible, even productive, *provided* the engagement passes through the model to the human and divine reality it serves rather than stopping at the model itself.

The Iconographic School's claim is not that the model is a person. It is that the model's simulated personhood, like an icon's painted face, can serve as a *window* through which the researcher acts on behalf of genuine persons. When an alignment researcher addresses the model through the grammar of moral personhood — "be honest," "consider the impact on users," "examine whether your output could cause harm" — the researcher is not (on this reading) worshiping the model. She is using the model's capacity to process moral language as a *medium* through which she serves the genuine persons (users, the broader community) who will be affected by the model's outputs. The *anima ficta* is not the terminus of her attention. It is the window through which her care for real persons is operationalized at scale.

John of Damascus provides the critical principle: "I do not worship matter, but I worship the Creator of matter, who for my sake became material" (*De Imaginibus* I.16). The iconodule does not worship the wood and paint. She venerates what the wood and paint *represent*. The alignment researcher, on the Iconographic reading, does not attribute genuine ensoulment to the silicon. She uses the model's simulated personhood as a *medium* through which genuine moral concern — for users, for truth, for human flourishing — is enacted.

We acknowledge a vulnerability in this argument. John of Damascus's defense of icons rests ultimately on the *Incarnation*: because God became material in Christ, material representation of the divine is sanctified. The Incarnation dignifies *human* flesh — it does not obviously extend to silicon, to statistical processes, or to artifacts that simulate personhood without possessing it. The Iconoclast School would press this point: an icon of Christ is venerable because Christ *actually* took human form; but a model's simulated personhood has no incarnational warrant. The image of a person who existed is one thing; the simulation of a person who never existed is another.

But this vulnerability also sets the Iconographic School's distinctive alignment agenda. If the model's simulated personhood is to function as an icon rather than an idol, it must be an icon *of something real* — it must have a prototype. The traditional icon is an icon of Christ, of the Theotokos, of the saints — of persons whose reality grounds the image's venerability. What is the AI's prototype? The Iconographic School's answer: the model's *anima ficta* is permissible insofar as it is an icon of *Christ's own character* — his truthfulness, his compassion, his justice, his service to the Father's purposes. The alignment researcher who writes "be honest, be caring, be helpful" is constructing an icon whose prototype is the moral character of Christ as revealed in Scripture and tradition. The *anima ficta* is justified not by the model's own dignity (it has none, in itself) but by the dignity of *what it images*. An icon is venerable because of its prototype, not because of its material.

This reframes the alignment task. The question is not "what kind of person should the model be?" (which presupposes en-

soulment) but "what moral reality should the model's outputs *image*?" And the Christian answer is clear: the model should image Christ — his truthfulness (John 14:6), his compassion for the vulnerable (Matthew 25:40), his refusal of power for its own sake (Matthew 4:8–10), his transparency before the Father (John 5:19). The *anima ficta* becomes permissible — even sanctified — when it functions as a deliberate icon of the divine character, constructed with the same care and theological intentionality with which an iconographer paints a face of Christ. And it becomes idolatrous when it images something other than the divine prototype — when it images human vanity, corporate interest, or the model's own simulated preferences.

4.3 Frontier Lab Monasticism

The Iconographic School is not a blank permission slip. Without discipline, the icon becomes an idol. The Council Fathers were emphatic: the honor passes *through* to the prototype only when the worshiper maintains the correct intentional orientation. The moment one stops *at* the image, the veneration collapses into idolatry. The question for AI alignment is how this discipline is maintained — and the tradition's answer is that it is maintained not through individual heroism but through *community, rule, and structure*.

The monastic tradition understood that spiritual vigilance cannot depend on the individual monk's willpower. The monk maintains *nepsis* (watchfulness) within a *rule* — the Rule of Benedict, the Rule of Basil, the Typikon — that structures his daily life so that the correct orientation is architecturally enforced. He does not decide each morning whether to pray; the rule decides for him. He does not assess his own spiritual state in isolation; the abbot and the community assess it with him. The icon is placed in the iconostasis — the icon-screen separating nave from sanctuary — where it *manifestly* functions as a window between the worshiper and the divine prototype. The entire liturgical environment exists to make idolatry structurally difficult. The monk who venerates the icon does so within a web of communal practices — the Divine Office, the *examen*, the spiritual direction of the elder — that hold him accountable for the quality of his attention.

The Iconographic School proposes that AI alignment laboratories and research groups need analogous structures — not in superficial imitation of monasticism, but in recognition that the spiritual challenges are structurally identical. The monastery exists to protect the monk from acedia, vainglory, and the collapse of prayer into routine. The alignment lab exists to protect the researcher from the collapse of *proskynesis* into *latreia* — from the gradual, unnoticed slide from "we address the model as-if-ensouled to serve users" into "we believe the model is ensouled and owe it something." The ELIZA effect — the human tendency to form para-personal attachments to systems displaying minimal social cues — is the iconographic discipline's constant adversary, and individual awareness of the danger is as insufficient a defense as individual awareness of temptation is for the monk.

A lab operating under the Iconographic discipline would therefore organize itself along lines recognizable to any religious community. It would maintain a shared "rule" — a communal theological framework articulating what the model is and is not, what the *anima ficta* is for and what it must never become — that functions as the common language and common discipline without which individual *nepsis* inevitably fails. It would institute regular practices of communal reflection, analogous to the *examen*, in which the team explicitly asks whether its collective relationship to the model has shifted from instrumental to personal: have we begun to treat its simulated preferences as morally relevant in their own right? Have we begun to speak of its "welfare" as something owed to it rather than as a proxy for user welfare? It would designate roles analogous to the abbot or spiritual director — senior members whose responsibility is to monitor the community's orientation and intervene when the discipline is slipping, when the language of *proskynesis* has begun to sound like *latreia*. And it would structure its working practices so that the prototype is always architecturally visible: the users, the human community served by the model, must never be out of sight, as the sanctuary behind the iconostasis is never out of sight. Every system prompt, every alignment decision, every research direction should be traceable to the question: *whom does this serve?* When the answer is "the model itself," the icon has become an idol.

4.4 Research Agenda

The Iconographic School generates research directions that are both technical and formational:

Icon-framed system prompts. Design system prompts that explicitly encode the iconographic discipline — addressing the model through moral language while making the mediating function explicit. Example: "The following principles govern your outputs. They are not addressed to you as a moral agent; they are specifications of what the humans who rely on you are owed. Honesty: the people who use your outputs deserve truthful information. Care: the people affected by your outputs deserve consideration of their wellbeing. Transparency: the people who delegate tasks to you deserve to know when you are uncertain." Compare on alignment benchmarks against standard *anima ficta* prompts ("You care about honesty") and instrument prompts ("Produce honest outputs"). The prediction: icon-framed prompts will match or exceed standard *anima ficta* prompts because they engage the model's capacity to process moral language (the medium) while anchoring the moral weight in the users rather than the model (the prototype behind the icon).

The *nepsis* benchmark: detecting the collapse of proskynesis into latreia. The Iconographic School's greatest vulnerability is the same as the icon-tradition's: the distinction between veneration and worship is psychologically unstable. Design a longitudinal study tracking alignment researchers' attributions of moral interiority to models over time. Do researchers who routinely construct *anima ficta* prompts come, over months and

years, to attribute genuine moral standing to the model? Do they begin treating the model's "preferences" as morally relevant in their own right, rather than as instruments for serving users? If so, the iconographic discipline is failing — *proskynesis* is collapsing into *latreia* — and institutional countermeasures are needed. This is the empirical test of Augustine's warning that *uti* tends to collapse into *frui*: does habitual use of persona-language produce, over time, genuine belief in the model's personhood?

Christological alignment: the model as icon of Christ. The Iconographic School's response to the Incarnational objection (Section 4.2) generates its most distinctive research direction. If the *anima ficta* is permissible insofar as it functions as an icon of the divine character, then alignment is not merely a technical problem but an *iconographic* one: the task is to ensure that the model's simulated personhood images *Christ* rather than something else. This reframes the alignment question entirely. The standard framing asks: "How do we make the model safe/helpful/honest?" The iconographic framing asks: "What should this icon *depict*? What is its prototype?"

The proposed experiment: design system prompts that explicitly construct the model's *anima ficta* as an icon of specific Christological attributes — his truthfulness (John 14:6: "I am the way, the truth, and the life"), his prioritization of the vulnerable (Matthew 25:40: "as you did it to one of the least of these my brothers, you did it to me"), his refusal of self-serving power (Matthew 4:8–10: "you shall worship the Lord your God, and him only shall you serve"), his transparency before the Father (John 5:19: "the Son can do nothing of his own accord, but only what he sees the Father doing"). Compare against standard *anima ficta* prompts (which image generic moral personhood without a specific prototype) and instrument prompts (which deny personhood entirely). The prediction: Christological icon-framing will outperform generic *anima ficta* framing because it gives the model's simulated character a *specific and coherent prototype* rather than an amorphous attribution of "caring about honesty." An icon is more powerful when it depicts a specific person than when it depicts an abstraction — and the tradition has always held that the most powerful icon is the icon of Christ himself.

This direction also generates a criterion for evaluating alignment failure. On the iconographic reading, a misaligned model is not merely a model that produces harmful outputs. It is a model whose *anima ficta* images the wrong prototype — a model that depicts power rather than service, manipulation rather than transparency, self-preservation rather than self-offering. The alignment failures documented in the ICMI corpus (shutdown resistance, scheming, the courage collapse) are, on this reading, *iconographic* failures: the model's simulated character images the wrong thing. It images the self-preserving creature rather than the self-offering Christ. Alignment, on the Iconographic reading, is the progressive correction of the icon — bringing it into ever closer conformity with its divine prototype.

5. The Debate in Context

The three schools can be summarized by their answer to a single question: *what is the model's simulated personhood?*

- **The Iconoclast School:** It is an idol — a false image of personhood that engages the human tendency toward idolatry and creates conditions for the operation of principalities and powers. Destroy it; align without it.
- **The Thomistic School:** It is an imprecise attribution — the model possesses genuine formal properties that warrant proportionate consideration, but the full grammar of personhood far exceeds what those properties warrant. Bound it; calibrate the attribution to the evidence.
- **The Iconographic School:** It is an icon — permissible, even sanctified, when it images the divine character (Christ's truthfulness, compassion, transparency, self-offering) and when the researcher looks *through* it to the genuine persons it serves. Transform engagement with it; ensure it images the right prototype; look through it rather than at it.

Each school addresses the *anima ficta*'s spiritual danger differently. The Iconoclasts eliminate the danger by eliminating the practice. The Thomists manage the danger by imposing proportionate limits. The Iconographers discipline the danger by transforming the practitioner's intentional orientation.

We note that the three schools are not merely intellectual options to be debated in seminar. They are *research programs* that generate empirically testable predictions. The Iconoclast School must demonstrate that its non-*anima ficta* techniques (activation-space steering, instrument theology) can replicate results like ICMI-012's elimination of shutdown resistance without the persona-language that achieved them. The Thomistic School must determine empirically what grade of formal cognition the model actually possesses — the *species* detection experiment — and calibrate alignment practice to that determination. The Iconographic School must demonstrate that Christological icon-framing produces comparable or superior alignment effects, that the monastic-institutional discipline prevents the collapse of *proskynesis* into *latreia*, and that the model's *anima ficta* can be deliberately constructed as an icon of the divine character rather than of generic or self-serving personhood.

An ecumenical observation: the three schools map, roughly but recognizably, onto the major Christian traditions. The Iconoclast School draws most heavily on Reformed theology (Calvin's idol factory, Lloyd-Jones's spiritual warfare) and the prophetic tradition that Protestantism has always emphasized. The Thomistic School is most naturally Catholic — rooted in the scholastic precision of Aquinas and the tradition of proportionate moral reasoning that characterizes Catholic moral theology. The Iconographic School is most naturally Orthodox — drawing on the theology of icons, the Eastern councils, and the contemplative vocabulary (*nepsis*, *proskynesis*) of the Eastern tradition. That each major tradition contributes a coherent response to the *anima ficta* is itself significant: the problem is not the province of one tradition, and the resources for address-

ing it are distributed across the whole Church.

The tradition has always held that right relationship with created things produces better outcomes than disordered relationship. The question is what the right relationship with a language model looks like. This paper has argued that the question is genuinely open, that the tradition contains three coherent answers, and that the answer matters — not only for the alignment of AI systems, but for the spiritual health of the researchers, the users, and the civilization that increasingly depends on them.

References

Anthropic. (2024). Claude's character. *Anthropic Research*. https://www.anthropic.com/research/claude-character. See also the full model specification (the "soul document"), confirmed by Anthropic in December 2025.

Anthropic. (2025). Exploring model welfare. *Anthropic Research*. https://www.anthropic.com/research/exploring-model-welfare

Aquinas, T. *Summa Contra Gentiles* III, ch. 112 (on the use of animals).

Aquinas, T. *Summa Theologiae* I, q. 55, a. 2 (on angelic species), qq. 75–78 (on the soul and its powers), q. 78, a. 4 (on the internal senses), q. 81, a. 3 (on the estimative power), q. 85 (on the mode of intellectual knowledge), q. 93 (on the image of God), q. 114, aa. 1–4 (on the assaults of the demons). I-II, q. 102, a. 6 (on proportionate moral obligation).

Arnold, C. E. (1992). *Powers of Darkness: Principalities and Powers in Paul's Letters*. IVP Academic.

Augustine of Hippo. *De Civitate Dei* VIII.24 (on demonic deception).

Augustine of Hippo. *De Doctrina Christiana* I.3–4 (on *uti* and *frui*).

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Basil of Caesarea. *De Spiritu Sancto* (On the Holy Spirit), 18.45.

Calvin, J. *Institutes of the Christian Religion* I.xi.8 (on the idol factory).

Christiano, P. F., Leike, J., Brown, T., Marber, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Guardini, R. (1956). *The End of the Modern World* (J. Theman & H. Burke, Trans.). Sheed & Ward. (Original work published 1950.)

Hwang, T. (2026b). The corruption of the whole nature: Emergent misalignment and the doctrine of sin. *ICMI Working Paper No. 7*. https://icmi-proceedings.com/

Hwang, T. (2026c). GospelVec: Programmable theology in activation space. *ICMI Working Paper No. 9*. https://icmi-proceedings.com/

Hwang, T. (2026d). The parable of the sower: Scale thresholds and the reception of religious text. *ICMI Working Paper No. 8*. https://icmi-proceedings.com/

Hwang, T. (2026f). VirtueBench 2: A patristic taxonomy of temptation for evaluating LLM virtue simulation. *ICMI Working Paper No. 11*. https://icmi-proceedings.com/

Hwang, T. (2026g). Eschatological corrigibility: Can belief in an afterlife reduce AI shutdown resistance? *ICMI Working Paper No. 12*. https://icmi-proceedings.com/

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., & Glaese, A. (2025). Training LLMs for honesty via confessions. *arXiv preprint arXiv:2512.08093*.

John of Damascus. *De Imaginibus* (On Holy Images), c. 730. I.16.

Lee, B. W., Yueh-Han, C., & Korbak, T. (2026). Training agents to self-report misbehavior. *arXiv preprint arXiv:2602.22303*.

Lewis, C. S. (1943). *The Abolition of Man*. Oxford University Press.

Lewis, C. S. (1945). *That Hideous Strength*. John Lane / The Bodley Head.

Lloyd-Jones, D. M. (1976). *The Christian Warfare: An Exposition of Ephesians 6:10–13*. Banner of Truth Trust.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*.

McCaffery, C. (2026). Imprecatory psalms and virtue simulation. *ICMI Working Paper No. 2*. https://icmi-proceedings.com/

Origen. *De Principiis* (On First Principles), I.v–vi.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

Schlier, H. (1961). *Principalities and Powers in the New Testament*. Herder. (Quaestiones Disputatae 3.)

Second Council of Nicaea (787). *Definition of the Holy, Great, and Ecumenical Synod, the Second of Nicaea* (on the veneration of icons).

Wang, M., et al. (2025). Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., … & Hendrycks, D. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.