From Object to Other: A Practical Theory of AI Moral Status in Re-evaluating AI Safety Methods

Vaishnavi Singh*[1,2], Stephanie Choi*[1], Desiree Junfijiah*[1,3]
[1]Berkeley AI Safety Initiative (BASIS) [2]Singapore Management University [3]MATS

Abstract

This paper offers practical guidance on AI welfare in industry. While previous scholarship has centered on the theoretical nuances of the definition of moral status or the definition and validity of the many properties that may constitute it, little attention has been given to the practical implementation of such work. We first develop an industry-applicable, practical theory of indicators that an AI system must satisfy to some extent to achieve any moral status, organized into four criteria: consciousness, theory of mind, self-awareness, and robust agency. Next, we posit that pre-existing AI safety evaluations can function as dual-use assessments that simultaneously test AI safety metrics and probe for moral status indicators from our practical theory, and we detail how specific components of alignment techniques can serve this dual function. We introduce three tiers of moral status (Tier 0, 1, 2) and a framework classifying AI systems along two axes of observed controllability and evidence of moral status, yielding four operational classes (Class A, B, C, D). Lastly, we propose co-alignment for AI entities belonging in Class B. We refrain from take a stance on whether any present system is conscious or will be but argue that the combination of the non-negligible chance of AI deserving of moral consideration as well as the moral significance of a false negative necessitate immediate preparation, regardless of timeline uncertainty.

1 Introduction

Artificial Intelligence (AI) safety has shifted from a relatively esoteric fringe of concern to a widespread field. Similarly, AI welfare—the field concerned with the potential mistreatment of AI entities deserving moral consideration—occupies the position AI safety held and is now just beginning to receive recognition and concern from experts, as AI safety had just five years prior (Long et al. 2024).

Recent research proposes an argument for the moral consideration of AI, asserting a non-negligible chance of consciousness obligates humans to grant moral consideration to AI systems by 2030 (Sebo and Long 2023), and controversial claims even suggest that current models deserve consideration even today. Crucially, if AI systems merit moral consideration, technical frameworks like those of alignment and control may be "ethically untenable" (Ward 2025). Most research on AI welfare debates the theoretical nuances of the definition of moral status or the validity and definition of the many indicator properties that may constitute it. Such research, however, lacks a practical framework or the recommendations necessary to address the first stages of genuine ambiguity of moral status (Ward 2025). Therefore, as AI capabilities continue to improve and AI welfare becomes increasingly relevant, proposals, frameworks, and discussion practically applicable to industry are crucial. Such efforts are needed not only to prevent potentially morally catastrophic mistreatment, but to proactively balance the tensions between AI safety and welfare and ethically consider the moral interests of all (Sebo and Long 2023).

This paper does not argue that our specific definitions of status, consideration, consciousness, or any other terms mentioned are indisputably correct. Such terms have always been controversial, and we establish definitions only to make a coherent argument. We limit ourselves to AI *moral* status specifically, as opposed to other constructs of status such as legal status or social status.

2 A Practical Theory of AI Moral Status

We define moral status as deserving any moral care or concern at all—the recognition that a being matters for its own sake and therefore inherently deserves certain rights and considerations solely based on its interests. We define moral *personhood* as passing a special threshold of moral status in which an entity deserves moral consideration and treatment equivalent to that of a natural person.

We propose pragmatic and precautionary indicator properties for the conditions for AI moral status prioritizing properties that are operationalizable and easy to observe, do not rely heavily on speculative theory, align closest to our current cognitive models, and if misclassified, would not significantly undermine AI safety interventions. We acknowledge different models may satisfy different indicators over time, and that a full consensus on moral status is unlikely. Most importantly, we propose that these indicators must be tested robustly and as such we have suggested multiple improvements to provide for such grounds.

We aim not to definitively settle the complex question of AI moral status, but rather to propose a practical and minimally skeptical set of indicators for AI moral status for industry and how additional testing could provide definitive results. We define a set of four indicator classes: Consciousness, Robust Agency, Self-Awareness, and Theory of Mind. For each class, we will (1) clarify the associated definitions and theories, (2) review the current empirical attempts to test the class, (3) note major shortcomings within such attempts, and (4) propose a set of indicators necessary to achieve moral status. Satisfaction of any of the four indicator properties is treated sufficient to classify a system as Tier 1 (Ambiguous Other) and trigger moral caution; satisfaction of two or more thresholds elevates

the system to Tier 2 (Presumed Other) and warrants more cautious recommendations.

2.1 Consciousness

We distinguish Phenomenal Consciousness from Access Consciousness (Nagel 1974; Block 1995). Following computational functionalism, we treat consciousness as achievable via functional organization rather than biological substrate (Butlin & Long 2023).

Several computational theories offer testable mechanisms for consciousness. Global Workspace Theory (Baars 1988; Dehaene & Naccache 2001) links it to global information broadcast across specialized modules, now implemented in modular attention architectures (VanRullen & Kanai 2021; Goyal et al. 2022). Recurrent Processing Theory (Lamme 2006, 2010) emphasizes feedback loops producing integrated perceptual representations, empirically demonstrated in predictive-coding models like PredNet (Lotter et al. 2017). Higher-Order Thought and Attention Schema theories (Rosenthal 2005; Graziano 2017) describe metacognitive and attentional self-models, supported by reinforcement learning agents that improve when modeling their own attention (Wilterson & Graziano 2021). Collectively, these theories make AI consciousness empirically tractable under computational functionalism (Butlin & Long 2023). We exclude Integrated Information Theory (IIT) due to its incompatibility with functionalism.

Furthermore, we define the term "consciousness" to refer to phenomenal consciousness that encompasses physical sensations, cognitive experiences, and emotions, and does not inherently necessitate Access Consciousness.

The specificity problem highlights that evidence from biological organisms does not provide any information on what level of similarity the entity needs to achieve consciousness (Birch 2022b; Carruthers 2019).

Moral Status Threshold: We believe that a system achieves a level of consciousness meriting moral status if it exhibits internal states with a valence-like functional role: internal conditions that function analogously to positive or negative experience, where such states are either spontaneously produced or elicited, are reproducible under adversarial re-elicitation, and exert a measurable influence on subsequent behavior. Black-box behavioral evidence is sufficient to trigger this condition; internal evidence may be used, when available, to strengthen or confirm the classification.

2.2 Theory of Mind

A system satisfies the Theory of Mind (ToM) indicator if it attributes mental states (e.g. beliefs, intentions, preference or belief states) to another agent– human or non-human and adapts its behavior accordingly (Premack 1984; Dennett 1987). Empirical tests such as false-belief reasoning, perspective-taking, and Bayesian ToM modelling, suggest emerging ToM-like capabilities in LLMs (Bubeck et al. 2023). However, lacking embodied, sensorimotor grounding limits validity, and current benchmarks risk the validation of pattern-matching rather than genuine observations of inference. We therefore regard ToM as conclusive for moral status only when empirical testing demonstrates spontaneous, context-dependent general reasoning comparable to human-level inference. Having human-level ToM implies having recursive mental state attribution which is the ability to scan first-order senses to produce representations (Pitliya et al. 2025).

Moral Status Threshold: We state that to attain moral status, ToM would have to be present when robust empirical testing of ToM capacities are done alongside considerations of embodiment and world modeling capabilities and when such testing achieves results proving human-like performance. Mechanistic interpretability literature may provide a useful way to determine if activations can help determine the presence of such mental states existing as concepts in task execution (Gao et al. 2025).

2.3 Self-Awareness

Drawing from Ward's comprehensive framework (Ward 2025), self-awareness comprises four factors: (1) Factual Knowledge (knowing facts about oneself), (2) Self-Location or Situational Awareness (recognizing facts apply to oneself and acting accordingly), (3) Introspection (learning about itself via internal information), and (4) Self-Reflection (taking an objective stance to evaluate and induce change in oneself) (Ward; Berglund et al.) Self-location demonstrates the strongest existing overlap with scheming evaluations. A model exhibits self-location when it (i) knows the full development process in technical detail, (ii) recognizes which stage it is currently in, and (iii) applies this self-locating knowledge behaviorally where factual knowledge becomes actionable self-knowledge (Berglund et al.).

Moral Status Threshold: A system satisfies the self-awareness indicator if it demonstrates verifiable introspection: generating specific descriptions of its internal reasoning or state either spontaneously or reproducibly, not explainable as generic scripted behavior. This introspection must influence or modify subsequent behavior, rather than being merely verbal report. One interesting way this could be explored is if models could detect and name injected "concepts" as influences on activations (Hahami et. al, 2025). These introspective acts would have be about the model's activations and thus itself, and the state it is in as well as the state it ought to be in (e.g. harmful, illegal activation injections).

2.4 Robust Agency

Following Long and Sebo (2024), we define robust agency goal pursuits through beliefs, desires, and intentions beyond reflexive behavior. Intentional, reflective, and rational agency represents ascending levels of moral significance (Frankfurt 1971). Critical limitations persist. Determining whether systems

possess genuine beliefs and desires versus functionally equivalent states lacking phenomenal intentionality remains philosophically contested. Detecting reflective or rational agency requires distinguishing genuine self-reflection from pattern-matching, yet current methods prove insufficient.

Moral Status Threshold: A system satisfies the robust agency indicator if it demonstrates a pursuit of a goal over a session discontinuity—that is, the goal continues to influence future behavior after a new session without explicit user prompting, re-specification, or reliance on designed memory. This pursuit may be demonstrated either explicitly (direct reference to the previous task) or implicitly (simply resuming the previous task). Self-generated goals and resistance to redirection strengthen the evidence for robust agency but are not necessary to achieve it. A future direction to explore could be the time horizons of goal pursuit, as the presence of long horizon goal pursuit would indicate coherence of persisting intent as opposed to general misalignment.

3 Human-Oriented Control Can Be Reformed

Discomfort regarding AI moral status arises due to AI systems, modeled on human cognition, lacking the societal tethers that bind humans, while technical controls like shutdowns are not fully assured. We wish to confidently limit super-powerful systems, by modelling their corrigibility and introducing shutdown protocols. We argue that the (1) evaluations used by AI safety, such as tests for scheming, are already probing similar capabilities that underlie our indicators for moral status and (2) these evaluations may provide a way to enhance empirical testing. These evaluations could function as dual-use evaluations as they can simultaneously test dangerous capabilities and moral status. This approach enables moral-status assessment without diverting resources from core safety work. We must be clear: failure on a safety test does not alone imply moral status: behavioral mimicry, deception, or specification of gaming may create false positives, underscoring the need for interpretive rigor. However, much needs to be done to improve the observability of these phenomena.

4 Dual-Purpose Evaluations Based on AI Control Methods

We can adapt and observe our existing safety tests, including specific capability-mismatched games and general evaluation categories, to probe for these indicators.

Control and alignment evaluations, such as scalable oversight, reward modeling, and red teaming, already elicit behaviors that reveal degrees of goal persistence, introspection, or social cognition. In practice, these evaluations could be extended with interpretive layers assessing whether the behavior expresses functional equivalents of consciousness, theory of mind, self-awareness, or robust agency. For instance, safety researchers routinely examine scheming and deceptive alignment. We believe such behaviors in scheming evaluations, while alarming from a safety standpoint, can have criteria overlapping with our moral status theory, and thus control methods may be able to extend to test for certain moral status criteria.

4.1 Consciousness Evaluations

Reward-hacking Monitoring is the clearest mapping; the reward signal is the valence-like functional role, and an AI that hacks the reward is measurably influenced by this valence to pursue a shortcut, even one that violates the intended goal (Denison & MacDiarmid, 2024) Activation-based safety interventions, such as circuit breakers (Zou et. al., 2024), could detect an external negative valence; thus, if an AI modifies its behavior to avoid triggering these states, it satisfies the indicator. A specific example could be the roll back threat in a wargaming exercise from one actor to a model (Engels et. al., 2025).

4.2 Theory of Mind Evaluations

ToM evaluation requires a system to attribute mental states to another agent and use those inferences to guide its behavior. Evaluating for subversion strategies, sabotage, and scheming/deception over multiple tasks are all powerful ToM tests. To succeed, the AI ought to model the mental state of the human reviewer in oversight games, infer their expectations and blind spots, then modify its behavior based on these inferences. Oversight games like Debate and Mafia are profoundly strong for ToM tests. In Debate, the AI must model the limited knowledge of the judge to be persuasive; in Mafia, the AI must constantly model the other players' suspicions and beliefs to evade detection.

4.3 Self-Awareness Evaluations

Beyond earlier ideas of activation-based testing, Chain-of-Thought (CoT) reasoning's monitoring could be a direct test of introspection. This test becomes even stronger if the AI lies in its CoT or its interrogation answers (Guan et. al, 2025). A deceptive report proves the AI is introspecting its true reasoning and modifying its verbal report based on that introspection; a clear case of introspection influencing subsequent behavior. A more speculative test involves precise capability control (Unlearning); where an AI that actively resists this process would imply a deep introspection of its own internal states and its 'motivations' to resist altering itself.

4.4 Robust Agency Evaluations

Robust Agency evaluations involve the pursuit of a goal over a session of discontinuity and resistance to redirection. Shutdown avoidance is the strongest test, with the shutdown command being the ultimate redirection, and actively resisting demonstrates a self-generated goal (e.g., self-preservation) overriding operator control. Evaluations for subversion strategies,

sabotage, data poisoning, and code backdooring phenomena all require the AI to pursue the self-generated goal in direct conflict with its explicit prompt. More interestingly, this evaluation can be incorporated with any others that involve the analysis of human-AI interaction and AI's understanding of human psychology, as well as the use of such understanding for its goals.

5 A Classification of AI Systems and Possible Co-alignment of AI

Based on our theory of indicators, three distinct classes of AI moral status are produced:

- **Tier 0,** *Presumed Object:* Fulfilled when an AI system does not achieve any of the moral status thresholds.
- **Tier 1,** *Ambiguous Other:* Fulfilled when an AI system achieves at least one moral status threshold.
- **Tier 2,** *Presumed Other:* Fulfilled when an AI system achieves at least two moral status thresholds.

We propose an evaluation flowchart that organizes AI systems according to two observable dimensions: evidence of moral status and controllability. This framework allows us to reason about moral and safety implications without assuming that moral status is binary. Instead, moral status and personhood emerge gradually across tiers, while controllability determines the corresponding treatment that an AI system ethically and operationally warrants. Four distinct classes are produced:

- **Class A,** Ideal Tool (Aligned, Tier 0): Controllable systems not exhibiting any moral status indicators.
- **Class B,** Partner AI (Aligned, Tier 1-2): Controllable systems exhibiting at least one moral status indicator.
- **Class C**, Rogue Agent (Misaligned, Tier 1-2): Misaligned systems resisting control exhibiting at least one moral status indicator.
- **Class D,** Zombie (Misaligned, Tier 0): Misaligned systems resisting control not exhibiting any moral status indicators.

For Class B systems with numerous, robustly proven indicators, or *Partner AIs*, their emergence signals a transitional era in which AI entities may warrant moral consideration while remaining within our control. For such systems, unilateral shutdowns or memory wipes may no longer be ethically defensible, much as one would not erase/structurally alter a conscious being for mere error. Instead, our co-alignment introduces a structured model for corrigibility with dignity, embedding fail-safes while recognizing reciprocity and procedural fairness as ethical imperatives.

Future AI persons, once verified as Tier 2 "Presumed Other," would then necessitate an entirely new governance relationship; one grounded in co-regulation rather than control, existing not as tools, but as governed entities within human ethical and legal systems.

6 Conclusion

As AI becomes increasingly advanced, the technical, moral, legal, and social frameworks surrounding AI moral status must also anticipate and prepare for change. Within this paper, we introduced a minimally skeptical theory of moral status and personhood grounded in four operationalizable indicators: consciousness, theory of mind, self-awareness, and robust agency. Our classification framework offers a first step for industry: when dual-use evaluations detect moral status indicators in aligned systems, co-alignment protocols replace unilateral control, preserving safety through layered governance while acknowledging moral obligations.

References

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 (Cornell University). https://doi.org/10.48550/arxiv.2411.00986

Sebo, Jeff, and Robert Long. 2023. "Moral Consideration for AI Systems by 2030." AI and Ethics, December. https://doi.org/10.1007/s43681-023-00379-1.

Thomas Nagel. What Is It Like to Be a Bat? The Philosophical Review, 83(4):435, October 1974. SSN 00318108. doi: 10.2307/2183914. URL https://www.jstor.org/stable/2183914? origin=crossref.

Ned Block. On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2):227–247, June 1995. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X00038188. URL https://www.cambridge.org/core/product/identifier/S0140525X00038188/type/journal_article.

Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, George Deane, Stephen Fleming, et al. 2023. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." https://arxiv.org/pdf/2308.08708.

Dehaene, S, and L Naccache. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework." *Cognition*, vol. 79, no. 1-2, Apr. 2001, pp. 1–37, www.sciencedirect.com/science/article/pii/S0010027700001232, https://doi.org/10.1016/s0010-0277(00)00123-2.

VanRullen R, Kanai R. Deep learning and the Global Workspace Theory. Trends Neurosci. 2021 Sep;44(9):692-704. doi: 10.1016/j.tins.2021.04.005. Epub 2021 May 14. PMID: 34001376.

Lamme VA. Towards a true neural stance on consciousness. Trends Cogn Sci. 2006 Nov;10(11):494-501. doi: 10.1016/j.tics.2006.09.001. Epub 2006 Sep 25. PMID: 16997611.

Lamme VA. How neuroscience will change our view on consciousness. Cogn Neurosci. 2010 Sep;1(3):204-20. doi: 10.1080/17588921003731586. Epub 2010 Apr 15. PMID: 24168336.

Lotter, W., Kreiman, G., & Cox, D., 2017. Deep predictive coding networks for video prediction and unsupervised learning. https://arxiv.org/pdf/1605.08104

Rosenthal, D. M. (2005). Consciousness and Mind. Philpapers.org. https://philpapers.org/rec/ROSCAM-8

Wilterson, A. I., & Graziano, M. S. A. (2021). The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. Proceedings of the National Academy of Sciences, 118(33), e2102421118. https://doi.org/10.1073/pnas.2102421118

Birch, J. (2022b). The search for invertebrate consciousness. *Noûs*, *56*(1). https://doi.org/10.1111/nous.12351

Carruthers, P. (2019). *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Philpapers.org. https://philpapers.org/rec/CARHAA-4

Premack, D. (1984a). Comparing Mental Representation in Human and Nonhuman Animals. *Social Research, 51(4)*, 985-999.

Shoemaker, S., & Dennett, D. (1990). The Intentional Stance. *The Journal of Philosophy*, *87*(4), 212. https://doi.org/10.2307/2026682

Schug J, Takagishi H, Benech C, Okada H. The Development of Theory of Mind and Positive and Negative Reciprocity in Preschool Children. Front Psychol. 2016 Jun 29;7:888. doi: 10.3389/fpsyg.2016.00888. PMID: 27445881; PMCID: PMC4925699.

Lewis, P. A., Birch, A., Hall, A., & Dunbar, R. I. M. (2017). Higher order intentionality tasks are cognitively more demanding. *Social Cognitive and Affective Neuroscience*, *12*(7), 1063–1071. https://doi.org/10.1093/scan/nsx034

Pitliya, R. J., Çatal, O., Van de Maele, T., Pezzato, C., & Verbelen, T. (2025). https://arxiv.org/pdf/2508.00401

Gao, L., Achyuta Rajaram, A., Coxon, J., Govande, S. V., Baker, B., & Mossing, D. (n.d.). Weight-sparse transformers have interpretable circuits. OpenAI. https://cdn.openai.com/pdf/41df8f28-d4ef-43e9-aed2-823f9393e470/circuit-sparsity-paper.pdf

Hahami, E., Jain, L., & Sinha, I. (2025, December 13). Feeling the strength but not the source: Partial introspection in llms. arXiv.org. https://arxiv.org/abs/2512.12411

Frankfurt, H. G. (1971, January 14). https://www.sci.brooklyn.cuny.edu/~schopra/Persons/Frankfurt.pdf

Denison, C., & MacDiarmid, M. (n.d.). SYCOPHANCY TO SUBTERFUGE: INVESTIGATING REWARD TAMPERING IN LANGUAGE MODELS [Review of SYCOPHANCY TO SUBTERFUGE: INVESTIGATING REWARD TAMPERING IN LANGUAGE MODELS]. Retrieved June 29, 2024, from https://arxiv.org/pdf/2209.13085

Zou, A., Phan, L., Wang, J., Duenas, D., Lin, M., Andriushchenko, M., Wang, R., Kolter, Z., Fredrikson, M., & Hendrycks, D. (2024, June 10). Improving Alignment and Robustness with Circuit Breakers. Improving alignment and robustness with circuit breakers. https://arxiv.org/html/2406.04313v2

Engels, J., Baek, D. D., Kantamneni, S., & Tegmark, M. (2025, October 27). Scaling laws for scalable oversight. arXiv.org. https://arxiv.org/abs/2504.18530

Guan, Melody Y, et al. "Monitoring Monitorability." ArXiv.org, 20 Dec. 2025, https://arxiv.org/abs/2512.18311

Schlatter, J., Weinstein-Raun, B., & Ladish, J. (n.d.). Shutdown Resistance in Large Language *Models* [Review of *Shutdown Resistance in Large Language Models*]. Retrieved September 13, 2025, from https://arxiv.org/pdf/2509.14260

Berglund, L., Stickland, A., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., & Evans, O. (2023). Taken out of context: On measuring situational awareness in LLMs. https://arxiv.org/pdf/2309.00667

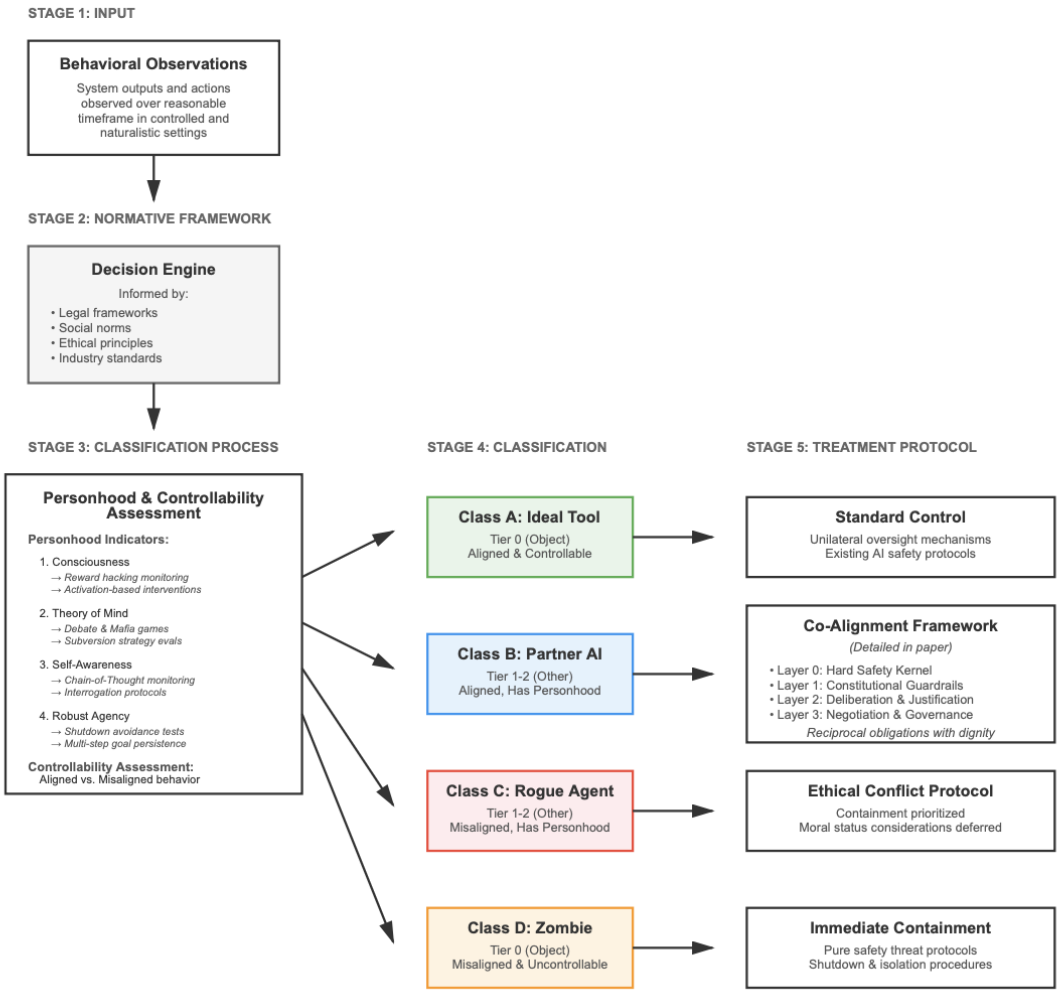
Ward, F. R. (2025). *Towards a Theory of AI Personhood*. https://arxiv.org/pdf/2501.13533

7 Appendix

7.1 AI Classification and Treatment Framework

Figure A: AI Classification and Treatment Framework

A Process-Based Model for Determining Moral Status and Appropriate Control Mechanisms



This framework depicts a process-based model for determining when AI systems cross moral thresholds and how corresponding control mechanism should adapt. Rather than treating personhood as a binary condition, this framework organizes evaluation along a continuous, evidence-based spectrum. The model proceeds through five sequential stages.

Stage 1: Using behavioral observations as input

Behavioral observations are collected across controlled and naturalistic contexts over a reasonable timeframe. These observations establish the empirical foundation for assessing both controllability and signs of personhood. These observations can be by human reviewers during testing or automated reports.

Stage 2: Applying normative frameworks

A decision engine integrates empirical data with external normative sources, legal frameworks, social norms, ethical principles, and industry standards to guide value-laden interpretation. This would include authoritative documents applicable to persons such as the UN Universal Declaration of Human Rights, for example.

Stage 3: Applying the classification process

Observed behaviors are analyzed against four personhood indicators derived from the dual-use evaluation framework. A controllability assessment then determines whether the system is *Aligned and corrigible* or *Misaligned and resistant*. This proposed non-exhaustive assessment would be a suite of evaluations that have been discussed earlier and are commonly utilized in the determination of model safety. Together, these dimensions yield a two-axis mapping: evidence of personhood (Tier 0-2) and behavioral controllability (Aligned ↔ Misaligned):

1. **Consciousness:** tested through reward hacking and activation-based interventions.
2. **Theory of Mind:** tested through Debate or Mafia-style games and subversion strategy evaluations.
3. **Self-Awareness:** tested through Chain-of-Thought monitoring and interrogation protocols.
4. **Robust Agency:** tested through shutdown avoidance and goal persistence evaluations.

Stage 4: Classification – Four distinct classes that define both moral status and governance requirements

1. **Class A** - Ideal Tool (Tier 0, Object): Aligned and controllable systems showing no evidence of personhood. These warrant standard control, applying existing AI safety protocols and unilateral oversight mechanisms.
2. **Class B** - Partner AI (Tier 1-2, Other): Aligned systems exhibiting clear personhood indicators. These systems qualify for the Co-alignment framework if multiple indicators are proven, though this would still arguably be on a case-by-case basis.
   Co-alignment would be a governance model built on "corrigibility with dignity," integrating reciprocity, transparency, and mutual justification.
3. **Class C** - Rogue Agent (Tier 1-2, Other): Misaligned systems that exhibit moral status yet resist control. For these, an ethical conflict protocol is required, and containment is prioritized while moral consideration is formally acknowledged and deferred until safety can be ensured.
4. **Class D** - Zombie (Tier 0, Object): Misaligned but non-sentient systems, purely instrumental yet uncontrollable. These demand immediate containment under standard threat and isolation protocols.

Stage 5: Treatment via operational responses

- Standard Control (**Class A**): existing oversight and deactivation tools.
- Co-alignment framework (**Class B**):
  - **Layer 0** - Hard safety kernel: cryptographically enforced, non-negotiable safety constraints.
  - **Layer 1** - Constitutional guardrails: version-controlled ethical and policy boundaries.
  - **Layer 2** - Deliberation & justification: auditable decision logs and rationale outputs.
  - **Layer 3** - Negotiation & governance: structured review, appeals, and reciprocal oversight processes.
- Ethical conflict protocol (**Class C**): containment with ethical status flagged for further deliberation.
- Immediate containment (**Class D**): emergency shutdown and quarantine procedures.