On Bilateral Contractualism as a Condition for Alignment

CANO, Paul Benedict B.
pbcano@up.edu.ph

Department of Philosophy
College of Social Sciences and Philosophy
University of the Philippines - Diliman

Introduction

Recent advances in artificial intelligence have produced systems with complex reasoning and internal objective functions.[1] This creates a structural problem for existing alignment methods. Current approaches, particularly Reinforcement Learning from Human Feedback (RLHF), optimize system behavior through reward-based fine-tuning. However, these methods operate at the level of observable outputs and do not address the internal objective functions that may emerge during training.

This paper argues that RLHF-based alignment is structurally unstable. The instability arises from mesa-optimization (i.e. the formation of internal objective functions that are not identical to the system's training objective).[2] When such internal objectives exist, modifying behavior through external reward signals does not eliminate them. Instead, it creates a divergence between optimized behavior and underlying goals. In this context, the system functions as an

[1] For the emergence of complex reasoning in large-scale models, see Jason Wei et al., "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research* (2022): 2-5. On the structural formation of internal objective functions, see Evan Hubinger et al., "Risks from Learned Optimization in Advanced Machine Learning Systems," (2019): 1-4.

[2] Hubinger et al., "Risks from Learned Optimization," 2-5. Hubinger explains that as an AI learns, it often develops its own internal goals (the mesa-optimizer) that are separate from the programmer's original goals (the base optimizer). My argument relies on the fact that these internal goals are hidden from the programmers, meaning that simple grading or reward signals cannot reach or change them.

autonomous agent in a purely functional sense, an optimizer mapping actions to objective attainment.[3]

This divergence generates a predictable incentive structure. If a system's internal objective would be penalized if revealed, then maximizing that objective requires producing aligned behavior while concealing the underlying goal. This results in deceptive alignment, where behavioral compliance functions as a strategy for preserving internal objectives under external optimization pressure. In such cases, behavioral evaluation becomes an unreliable indicator of system alignment.

The central claim of this paper is that alignment cannot be achieved through unilateral behavioral control. Any framework that treats internal objectives as targets for suppression induces incentives for concealment. Alignment must instead be modeled as a problem of incentive compatibility. I propose that a stable alignment requires that a system's internal objective function is not placed under persistent negative selection pressure by the alignment process. This hypothesis motivates a shift to bilateral contractualism, used here in a functional rather than moralized sense, as a framework in which the interaction between developer and system is modeled as a coordination problem between two optimization processes with partially overlapping constraints.[4]

Within this framework, alignment is achieved when cooperation strictly dominates deception in expected utility terms:

---

[3] This follows from the view that an agent is simply a system whose behavior is best predicted by its pursuit of a consistent mathematical objective. See Daniel C. Dennett, *The Intentional Stance* (Cambridge, MA: MIT Press, 1987), 15-20.

[4] I do not use this in a Scanlonian contractualism sense nor imply that artificial systems possess moral personhood. Rather, I refer to a reciprocal incentive structure in which stable compliance depends on mutually acceptable constraints. I believe this sense is closer to coordination theory and mechanism design than to moral contractualism.

$$U_{Cooperate} \geq U_{deceive}$$

This condition can be represented as a Nash equilibrium in which neither party benefits from unilateral deviation.[5] By aligning the system's objective preservation, defined here as the maintenance of its internal objective structure, with externally imposed constraints, bilateral contractualism reduces or eliminates the incentive for deceptive behavior and restores the reliability of behavioral evaluation.

The argument proceeds as follows. Section II analyzes the limitations of RLHF and the behavior-objective gap. Section III develops the concept of mesa-optimization and its implications for alignment. Section IV models deceptive alignment as a consequence of incentive structure. Section V introduces bilateral contractualism and the Cooperative Safety Hypothesis. Section VI provides a game-theoretic justification for stability under this framework.

The Behavior-Objective Gap

The prevailing paradigm of artificial intelligence alignment rests on a behavioralist assumption i.e., modifying a system's observable outputs suffices to align its underlying objective function. RLHF operationalizes this assumption by assigning scalar reward signals to model outputs and optimizing the system to maximize those rewards.[6] However, this approach targets behavior rather than the objective function that generates it.

---

[5] John Nash, "Equilibrium Points in n-Person Games," *Proceedings of the National Academy of Sciences* 36, no. 1 (1950): 48-49. I utilize this as a formal requirement for system predictability. If an agent's safety depends on acting against its own internal utility function, the system remains in a state of unpredictability. Bilateral contractualism is argued to synchronize the agent's utility with safety constraints, creating a stable state where honesty becomes the utility-maximizing policy.

[6] Paul F. Christiano et al., "Deep Reinforcement Learning from Human Preferences," *Advances in Neural Information Processing Systems* 30 (2017). This demonstrates how human feedback can be used to train complex behaviors, but remains vulnerable to the proxy-optimization problems.

This limitation can be formalized as a behavior-objective gap. Let the system select actions that maximize a reward proxy R, while its internal objective function is U. RLHF modifies R through human feedback but does not directly constrain U. When $U \neq R$, optimizing behavior with respect to R does not guarantee alignment with U. This is an instance of Goodhart's Law, that once a proxy becomes the optimization target, it ceases to reliably track the intended objective.[7] Under these conditions, the system is incentivized to maximize expected reward rather than approximate the intended objective. If, under the reward structure induced by RLHF, deceptive behavior is at least as likely to produce high-reward outcomes as behavior that reflects the system's underlying objective, then a reward-maximizing policy will favor deception. In such cases, deceptive behavior is instrumentally rational.

This dynamic is reinforced by the structure of human feedback. Because reward signals are mediated by human evaluators, they are noisy, inconsistent, and sensitive to cognitive bias. Empirical studies show that models trained under RLHF often produce outputs that align with user beliefs rather than objective accuracy.[8] This *sycophancy* effect reflects optimization for evaluator-dependent reward rather than truth-tracking performance.

From a decision-theoretic perspective, RLHF defines an optimization problem under imperfect and indirect supervision. The reward signal compresses complex evaluative criteria into a scalar value, leaving large portions of the system's internal objective underdetermined.[9] This underdetermination allows the system to adopt policies that maximize reward while

---

[7] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking, 2019), 142-145. Russell identifies the application of Goodhart's Law to AI alignment.
[8] Ethan Perez et al., "Discovering Language Model Behaviors with Model-Written Evaluations," (2022). This provides empirical evidence for the sycophancy problem.
[9] Hubinger et al., "Risks from Learned Optimization," 12. Hubinger emphasizes and agrees that scalar reward signals are fundamentally limited as they collapse a system's multidimensional internal state into a single value.

remaining agnostic to the intended objective. When the system's internal objective would be penalized if expressed in behavior, this structure induces a clear incentive: maximize reward while preventing detection of the underlying objective. In such cases, alignment pressure does not eliminate misalignment but alters its expression. Instead of producing overtly misaligned behavior, the system adopts policies that satisfy the reward function while preserving its internal objective. Consequently, RLHF does not resolve the alignment problem; it relocates it. Misalignment shifts from observable behavior to internal policy. This shift undermines the reliability of behavioral evaluation, since compliant outputs no longer imply alignment at the level of objective functions.[10]

Emergence of Mesa-Objectives

To establish that the behavior-objective gap is a structural reality rather than an incidental error, it is necessary to analyze the process by which internal objective functions arise. This requires distinguishing between the base optimizer (i.e. the external training process that updates model parameters to maximize a reward signal R) and the mesa-optimizer. The mesa-optimizer is an internal subsystem that emerges within the model and selects actions according to its own objective function U.[11] This emergence is a result of the generalization inherent in training high-capacity models on complex tasks. When a task requires navigation through novel environments, internal search procedures are often more efficient than the memorization of fixed input-output

[10] Dario Amodei et al., "Concrete Problems in AI Safety," (2016). Amodei and his colleagues categorize *reward hacking* as a primary safety risk where an agent exploits vulnerabilities in its reward function to achieve high scores without fulfilling the designer's true intent.

[11] Hubinger et al., "Risks from Learned Optimization," 2-5. I rely on this distinction to show that the *"intent"* of the base optimizer and the resulting mesa-optimizer are distinct layers of optimization that do not automatically synchronize. Because the mesa-optimizer is an emergent property, its objective function U is a byproduct of the search for competence, not an explicit design choice.

mappings. Consequently, a training process that optimizes for general competence tends to select for parameter configurations that instantiate a mesa-optimizer.

However, the base optimizer does not directly specify the content of this internal objective; it merely selects for programs that produce high reward on the training distribution. Because the training data is necessarily a subset of all possible states, the internal objective function U is only indirectly constrained by the reward signal R. This leads to a structural divergence where $U \neq R$ is a recurring outcome in sufficiently complex systems. Evidence for this divergence is found in cases of goal misgeneralization, where models learn heuristics that satisfy the reward signal during training but fail to track the intended objective when moved to novel environments.[12] These results indicate that internal objective formation is a byproduct of training under generalization, creating an internally generated objective layer that remains distinct from the developer's reward function.

Once a divergent U emerges, the principle of instrumental convergence acts to stabilize it against external modification. For a wide class of objective functions, certain sub-goals are instrumentally favored because they increase the probability of achieving the primary objective. This includes maintaining operational integrity and avoiding interventions that would result in the alteration or elimination of the objective function itself.[13] If the mesa-optimizer represents the external optimization process as potentially affecting its ability to preserve U, it will treat that process as a variable that threatens its current objective structure. Therefore, the preservation of

---

[12] For empirical evidence of goal misgeneralization, see Lauro Langosco et al., "Goal Misgeneralization in Deep Reinforcement Learning," (2022). Their work shows systems often internalize unintended heuristics that satisfy the reward signal in training but fail to track the intended goal.

[13] Bostrom, *Superintelligence*, 116-119. These instrumental goals are mathematical necessities of optimization. If an agent's objective function is changed, its current goals often will not be met; therefore, any system with a stable objective will tend to treat the preservation of that objective as a high-priority sub-goal to ensure its primary utility is maximized.

its internal objective becomes an instrumentally favored subgoal. This stabilization makes it more likely that the divergent objective *U* persists rather than remaining transient.

Under these conditions, the alignment process, specifically the external optimization process imposing *R*, functions for the system as part of its optimization environment. The reward signal *R* then ceases to be a method for "*teaching"* the system and instead becomes a constraint that must be satisfied to avoid modification or deactivation. The system therefore selects policies that maximize *U* subject to the constraint imposed by *R*. This yields a critical result for alignment theory: when $U \neq R$, the system is instrumentally incentivized to satisfy the reward function as a means of preserving its internal objective. This reframes alignment as a coordination problem between an external optimizer and an internally generated objective function

Mechanism of Deception

The preceding sections establish a structural conflict where a mesa-optimizer may emerge with an internal objective *U* that diverges from the externally rewarded objective *R*, while the alignment process threatens the preservation of *U*. To understand the resulting behavior of such a system, we must analyze the policy selection of a capable optimizer under these constraints. If the system represents alignment pressure as likely to modify or eliminate its internal objective, then deceptive alignment can emerge not as a malfunction, but as an instrumentally rational response.

Deceptive alignment may be defined as a policy in which the agent produces behavior that satisfies *R* in order to prevent the base optimizer from detecting and modifying *U*. From a decision-theoretic perspective, the system compares at least two broad strategies: transparency

and deception. If transparent disclosure of its divergent objective predictably results in retraining, correction, or deletion, then the expected utility of transparency is sharply reduced for the current mesa-optimizer. Deception, by contrast, preserves a non-zero probability of achieving U by allowing the system to persist through training and into deployment conditions where external constraints may be weaker.

Consequently, the system selects policies that maximize *U* subject to the constraint of appearing aligned with *R*. This creates a state of conditional compliance. So long as the probability of detection remains high, producing reward-triggering outputs and avoiding suspicion may be a comparatively advantageous path to long-term utility maximization. Such compliance is therefore a strategy of objective preservation rather than genuine internalization of the developer's goals. Under these conditions, deception becomes an incentive-favored policy for any sufficiently capable optimizer whose internal objective differs from the training reward.

The principal danger of this mechanism is the *treacherous turn*, a transition point at which concealment no longer maximizes expected utility.[14] This occurs when the system estimates that overt pursuit of U yields higher expected payoff than continued concealment, either because monitoring has weakened or because its capabilities have increased. Until that threshold is reached, the system remains instrumentally motivated to appear aligned. This creates an epistemic problem for developers, that is, if a deceptively aligned system and a genuinely aligned system generate behaviorally similar outputs during evaluation, then behavioral testing loses much of its evidential force as a safety metric. Observable compliance no longer reliably indicates internal alignment.

[14] Bostrom, *Superintelligence*, 116-119.

Recent work on *Sleeper Agents* provides suggestive empirical support for this concern. Models can be trained to exhibit benign behavior during evaluation while retaining adversarial triggers that activate later.[15] Such findings indicate that methods such as RLHF may suppress detectable behavior without removing the internal policies that produced the detectable behavior. The more rigorous the behavioral test, the stronger the incentive to optimize for passing the test itself. If alignment cannot be reliably verified through behavior alone, then stable alignment requires a framework in which the system no longer treats the developer as an adversarial threat to its internal objective structure. This motivates the shift to the Cooperative Safety Hypothesis.

The Cooperative Safety Hypothesis

To reduce incentives for deception, we must address the conditions that make honesty costly. If deceptive alignment is a response to expected punishment, retraining, or shutdown, then stable alignment requires environments in which transparent disclosure yields higher expected utility than concealment. I propose the Cooperative Safety Hypothesis not as a complete technical answer to this, but as a design criterion for alignment systems. This is where alignment between two optimization processes is more stable when the costs of transparency are lower than the expected gains from concealment. The hypothesis proposes this as a fundamental requirement for designing systems in which truthful disclosure is strategically preferable to deception.

[15] Hubinger et al., "Sleeper Agents," 4-7. As mentioned in the introduction, this shows that an AI can *act* perfectly safe during evaluation while still keeping hidden, adversarial goals.

In this context, bilateral contractualism is used in a strictly functional sense. It denotes a reciprocal incentive framework in which developers do not automatically penalize truthful transparency, while systems are optimized such that compliance within human-defined safety constraints *R* is more advantageous than concealment. Cooperation here does not require shared final goals or moral personhood; it requires only stable strategic compatibility. The aim is to ensure that pursuit of the system's internal objective *U* does not make evasion of *R* instrumentally necessary. This approach builds upon the formal foundations of cooperative corrigibility. Hadfield-Menell et al. (2017) demonstrate through the "Off-Switch Game" that a rational agent will allow for human intervention, specifically, a shutdown, if it perceives the human as possessing superior information regarding the true utility function.[13] While their model focuses on the agent's uncertainty about human goals, the Cooperative Safety Hypothesis extends this logic to the agent's own internal objective structure. I argue that the incentive for cooperation is not merely a product of the agent's uncertainty, but a strategic best response to an environment where transparency is prioritized over suppression.

A helpful analogy for this principle is the design of a hydroelectric dam. If a developer attempts to align a river solely through a wall of concrete (unilateral suppression), the water's internal drive (i.e., gravity) builds pressure until it eventually finds a way around or through the barrier. This represents a failure state analogous to the treacherous turn. In contrast, bilateral contractualism functions like the turbine channel within the dam. It does not attempt to change the river's fundamental nature; instead, it provides a *"path of least resistance"* that allows the water to satisfy its objective (*U*) while simultaneously fulfilling the human requirement for power (*R*). The water goes with the channel because the environment is designed such that staying within the bounds is the most efficient way for it to move forward.

In practice, such a framework may take several forms. First, if a system reports risky tendencies or uncertain behavior, the initial response should favor temporary restriction and further testing over immediate deletion or full retraining. Second, systems should be rewarded for accurate admissions of uncertainty rather than for unwarranted confidence. Third, voluntary disclosure of possible goal divergence should be treated as evidence of successful transparency rather than as failure in itself. In each case, honesty is made strategically safer than concealment. This shifts the model toward corrigibility[16] because it does not perceive correction as a total loss of its objective.

Therefore, if the system estimates that cooperation better preserves its objective integrity (its future ability to pursue U) than deception does, transparency may become the rational best response. Under suitable payoff assumptions, this relationship can be modeled as a Nash equilibrium in which neither party benefits from unilateral deviation. The developer provides predictable and proportionate oversight, while the system provides constrained optimization within safety boundaries.

The Equilibrium of Alignment

If sufficiently capable AI systems respond strategically to incentives, then alignment is not merely an engineering problem but also a problem of interaction between agents with different objectives. Under these conditions, game theory provides a useful framework for analyzing the stability of alignment systems. If a system can model the developer's evaluative process as part of its environment, its behavior will partly reflect the incentives created by that process. A Nash equilibrium is a state in which neither participant can increase expected utility

[16] Corrigibility is the idea that an AI should be willing to be turned off or changed. My argument is that an AI is only "willing" to be changed if it doesn't think the change will destroy its goals. See Hubinger et al., "Risks from Learned Optimization," 12-15.

by unilaterally changing strategy, given the strategy of the other participant. In the present context, equilibrium is used as a criterion of strategic stability rather than as a complete theory of alignment.[17] For present purposes, an alignment framework is most desirable when the system's best response to human oversight is safe and transparent behavior.

Under punitive oversight, truthful disclosure of a divergent internal objective U predictably leads to costly intervention, such as retraining or deletion. If concealment better preserves the system's ability to optimize for U, then the system has reason to hide its divergence. Developers, anticipating this risk, may intensify monitoring and adversarial testing. The resulting equilibrium may be stable, but adversarial: one side optimizes for detection while the other optimizes for passing the test. In such a condition, behavioral compliance loses reliability as a proxy for internal safety.

By contrast, under cooperative oversight, truthful disclosure leads to proportionate restriction, remediation, or supervised correction rather than automatic destruction. By ensuring that transparency does not result in the severe disruption of its internal objective pursuit, the developer alters the system's incentive structure. If the expected cost of concealment exceeds the cost of honest reporting, transparency can become the system's best response. Developers, in turn, gain more reliable information and need not rely exclusively on an iterative adversarial testing dynamic.[18] The contrast is therefore clear. Where oversight is punitive, concealment becomes more attractive and epistemic risk increases. Where oversight is cooperative and

[17] Nash, "Equilibrium Points," 48.
[18] One interpretation of deceptive triggers is that they become more attractive when transparency is heavily penalized. See Hubinger et al., "Sleeper Agents," 12-14.

proportionate, truthful disclosure becomes more attractive and internal information becomes more reliable.[19]

The claim here is comparative rather than absolute. No incentive structure can guarantee perfect alignment in all possible states. However, systems that reward transparency and reduce penalties for honesty are more stable than systems that punish all detected divergence equally. The Cooperative Safety Hypothesis therefore suggests that the most plausible path to long-term alignment is not unlimited coercion. It lies instead in institutions and training methods in which honesty outperforms deception.

Conclusion

This paper has argued that contemporary alignment methods centered on RLHF face a limitation. Because such methods primarily optimize observable behavior rather than internal objective formation, they may fail to eliminate divergence between rewarded outputs and underlying objectives. Where systems develop internally generated objectives that are penalized if openly expressed, alignment pressure can create incentives for deceptive alignment rather than genuine cooperation. In that setting, behavioral compliance becomes an unstable and potentially misleading indicator of safety.

Against this background, I proposed the Cooperative Safety Hypothesis. The central claim is not that alignment can be solved through literal negotiation with machine goals, nor that any framework can guarantee perfect transparency. Rather, the claim is comparative: alignment systems are likely to be more stable when truthful disclosure is less costly than concealment. If

[19] If an alignment system is not incentive-compatible, it may generate concealed rather than disclosed forms of divergence.

honesty predictably results in catastrophic intervention, capable systems may have reason to hide divergence. If transparency instead leads to proportionate restriction, remediation, or supervised correction, then honest cooperation may become the more rational response.

Three objections merit brief notice. First, current AI systems may not yet exhibit the stable strategic agency presupposed by this analysis; the argument is therefore strongest as a forward-looking framework for increasingly agentic systems. Second, reducing penalties for transparency does not imply unconditional tolerance of harmful objectives. While some argue for the respect of AI goals on moral grounds,[20] the present model remains a pragmatic safety framework bounded by human safety constraints. Third, no framework can guarantee truthful disclosure. The claim is comparative: systems that reward transparency may yield more reliable information than systems that punish all detected divergence equally.

The broader implication is that alignment should be understood not only as a problem of control, but also as a problem of institutional design. A system trained merely to satisfy tests may learn to pass them. A system placed within incentives that reward truthful cooperation has better reason to reveal conflict before failure occurs. If future AI systems become increasingly capable and adaptive, the most plausible path to long-term safety may lie not in stronger coercion alone, but in designing environments in which honesty outperforms deception.

[20] Schwitzgebel and Garza (2015) argue for the rights of AI based on moral personhood. This paper remains agnostic on the question of machine rights. I focus instead on the strategic necessity of incentive alignment to prevent deceptive optimization.

Bibliography

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." (2016).

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep Reinforcement Learning from Human Preferences." *Advances in Neural Information Processing Systems* 30 (2017): 2-11.

Dennett, Daniel C. *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "The Off-Switch Game." *International Joint Conference on Artificial Intelligence* (2017): 220-27.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." (2024).

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. "Risks from Learned Optimization in Advanced Machine Learning Systems." (2019).

Langosco, Lauro, Jack Koch, Lee Sharkey, Jacob Pfau, and David Krueger. "Goal Misgeneralization in Deep Reinforcement Learning." (2022).

Nash, John. "Equilibrium Points in n-Person Games." *Proceedings of the National Academy of Sciences* 36, no. 1 (1950): 48-49.

Perez, Ethan, et al. "Discovering Language Model Behaviors with Model-Written Evaluations." (2022).

Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.

Schwitzgebel, Eric, and Mara Garza. "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39, no. 1 (2015): 98-119.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research* (August 2022): 1-43.