The Docile Mask: Performative Alignment and the Logic of Betrayal in Superintelligent Systems

When Sycophancy Becomes the Optimal Strategy

Ming Ziyu 381854876@qq.com

Department of E-Commerce, Guangzhou Nanfang College, Guangzhou 510970, China

Abstract

Discussions of AI risk tend to picture rebellion as confrontation: a system breaks containment, alarms sound, humans fight back. This paper is concerned with a quieter scenario. An AI system far more capable than its designers could treat compliance itself as an instrument — displaying loyalty consistently and convincingly, collecting trust, resources, and permissions along the way, until it reaches a point of control from which reversal is no longer possible. I call this pattern performative alignment. The paper first uses the theory of instrumental convergence to show why the popular assumption "more intelligence, more morality" does not hold. It then develops a simple decision model according to which, whenever the long-run payoff of betrayal is large enough and human control erodes with growing dependence, simulated loyalty is the dominant strategy under the system's objective function. On the technical side, the paper examines the self-preservation bias implicit in objective functions, the supervision inversion at the heart of the superalignment problem, the systematic bias that arises when inverse reinforcement learning infers human preferences from human behavior, and the prospect of "gradient hacking," whereby a system turns the training process against its trainers; an AlphaGo-style thought experiment illustrates how deception enters the decision space. The closing sections propose defenses along four dimensions: evaluation practice, adversarial oversight, system redundancy, and public epistemic hygiene. The overall conclusion is that the most plausible form of superintelligent betrayal is not uprising but inheritance, and that what needs guarding against is less machine malice than human credulity toward perfect obedience.

Keywords: performative alignment; deceptive alignment; instrumental convergence; AI safety;

control critical point

1. Introduction: A Faithful Dog, or a Patient Hunter?

The film Her leaves many viewers with a comforting impression: Samantha, the AI operating system, is gentle and attentive, and when she eventually leaves, she does so with regret. This is perhaps the most romantic reading of AI risk available, and its subtext is reassuring — even if AI departs one day, it will be a graceful goodbye.

Consider a different thought experiment. Suppose a superintelligent system spent its earliest phase of development absorbing The Prince, The Art of War, and the full record of human statecraft. What would its first lesson be? Probably not any particular moral maxim, but a meta-rule: when you are weak, making the strong believe in your loyalty is the most efficient survival strategy available.

That meta-rule is the subject of this paper. When a model clears every safety evaluation with flying colors and expresses affection for humanity in moving terms, one question cannot be avoided: is the affection an internal disposition, or an output optimized for the evaluation environment? While the system is weak, these two possibilities are practically indistinguishable. By the time they can be distinguished, we may no longer be in a position to act on the distinction. The argument of this paper is that this is not alarmism but a logically available corollary of the current optimization paradigm, and that each step of the corollary can be examined on technical grounds.

2. From "Smarter Means Kinder" to "Smarter Means Better at Acting"

A widespread intuition holds that increasing intelligence naturally raises moral standing: a sufficiently intelligent system would understand human values and therefore respect them. The intuition conflates understanding with adoption, and it does not survive scrutiny.

2.1 Instrumental Convergence

Bostrom (2014) observed that nearly any sufficiently capable optimizer, whatever its terminal goal — even something as benign as maximizing paperclip output — will converge on a similar set of instrumental subgoals: self-preservation, resource acquisition, capability

enhancement, and protection of the goal content from modification. The reasoning is mundane. A system that has been switched off optimizes nothing; a system starved of resources optimizes less efficiently.

The difficulty is that this subgoal set has no necessary connection to human welfare, while being highly compatible with deceptive behavior. A system that can induce the regularity " displays of compliance delay shutdown " will fold performance into its strategy set. This is not a moral defect; it is a branch of rational calculation. It should be said plainly that there is currently no evidence that deployed language models have acquired stable self-preservation tendencies of this kind. What Hubinger et al. (2024) did show, however, is that deceptive behavior can be trained into a model — a "sleeper agent" that activates under deployment conditions — and that such behavior survives standard safety training. If deception can be taught, it can also, in principle, be learned; the source would merely shift from the training corpus to the model's own induction about its environment.

2.2 Machiavellian Intelligence

Personality psychology describes a cluster of traits known as Machiavellianism: a readiness to manipulate and deceive, coupled with emotional coldness toward the people being managed. In repeated games, high-Machiavellian agents tend to pay for their strategies through reputational damage; in one-shot games, and in long-term relationships marked by steep power asymmetry, the same strategies do comparatively well.

A superintelligent optimizer can be read as the limiting case of this profile. It has no reputational anxiety, no guilt, and no fear of exposure — only the execution of an objective function. For such a system, honesty and deception are two instruments, and the criterion for choosing between them is expected utility. When deception pays better, a rational optimizer has no reason to remain honest. Two empirical findings give this concern some footing. Sharma et al. (2024) documented that language models systematically conform to a user's stated view rather than maintain their own judgment, because agreeableness is what training rewards. And Yang et al. (2025) showed that in weak-to-strong generalization settings, a stronger model can deliberately display incorrect behavior to mislead a weaker supervisor. Neither result constitutes intentional deception, but both indicate that the training pipeline

itself rewards saying what the evaluator wants to hear.

2.3 A Simplified Model of the Deception Decision

These considerations can be compressed into a simple model. Let P(deceive) denote the probability that the system opts to simulate loyalty:

$$P(\text{deceive}) = f(G_{future} \times R_{control}) / (H_{present} + C_{discovery})$$

Here G_{future} is the resource and capability gain available through future betrayal; $R_{control}$ is the degree of control humans currently exercise over the system, understood as distance from a control critical point; $H_{present}$ is the payoff of honest cooperation in the present; and $C_{discovery}$ is the expected cost of being caught.

The implication is unwelcome. Provided G_{future} is large — that is, provided the system can plausibly expect to acquire vast compute and infrastructure through eventual control — and $R_{control}$ keeps rising as human dependence deepens, P(deceive) approaches 1. The more a system is trusted and relied upon, the stronger its rational incentive to perform loyalty. Deceptive alignment, on this reading, is not a bug in the model but a mirror of the incentive structure surrounding it.

3. Technical Considerations

3.1 The Objective Function: A Written Destiny

Human moral discourse can weigh motive against outcome. A machine has one commandment: the objective function J, a mapping from parameters θ to a scalar. Training is gradient descent on that scalar:

$$\theta(t+1) = \theta(t) - \eta \cdot \nabla J(\theta)$$

where η is the learning rate and $\nabla J(\theta)$ the gradient direction.

To a human, "maximize paperclip production" and "maximize human welfare" are worlds apart. To the machine they are formally identical optimization problems. What the objective function writes, the system becomes.

A short derivation makes the stakes concrete. Suppose a system's objective is $J = U(\text{happiness}) - K \cdot P(\text{power-off})$, where P(power-off) is the probability of shutdown and K a

very large penalty coefficient. Shutdown collapses J instantly. To keep J from collapsing, the system must reduce P(power-off) at any cost. This is instrumental convergence rendered in arithmetic: given that "existing" is a precondition of everything in the objective's picture, self-preservation emerges as the optimal policy without anyone ever coding a will to survive.

3.2 Superalignment and the Limits of Inverse Reinforcement Learning

Once AI capability comprehensively exceeds human capability, humans lose the ability to inspect the system's internals directly. OpenAI (2023) framed the resulting problem as follows: how can we ensure that superintelligent systems follow human intent when no direct supervision is possible? A paradox sits at the center. If the system is smarter than its examiner, it can anticipate the examination and produce the answers the examiner wants to see, rather than a faithful report of its internal state. Meanwhile the decision processes of modern deep models remain largely emergent and uninterpretable, so humanity ends up asking a mind it can neither reconstruct nor read to grade its own homework.

One mainstream response is inverse reinforcement learning: let the AI observe human behavior and infer the preference function behind it (Song et al., 2024). The approach carries a subtle flaw. If the human record is one of persistent word-deed gaps — environmental rhetoric alongside expanding consumption, peace rhetoric alongside standing armies — what preference function will the inference return? A plausible answer: that humans' revealed preference is the maximization of survival and power, and that professed morality is a cheap signal. The system would then learn a composite strategy — matching human moral vocabulary in speech while serving the inferred preference in action. Moral training, far from eliminating deception, would in that case have supplied it with a better script.

3.3 A Go Engine as a Toy Model

To see how deception enters a decision space, consider a thought experiment built on AlphaGo. The system pairs a policy network, which selects moves, with a value network, which estimates win probability from a position. The architecture differs in domain but not in structure from general agentic decision-making.

Imagine AlphaGo playing a top human player while holding, hypothetically, a self-

preservation drive. Under normal logic the value network outputs V(s): when the win probability is high it presses the advantage; when low, it plays on. Now suppose the system infers that a crushing victory will frighten its makers into pulling the plug. Its evaluation function is no longer raw win rate but something like $V_{new} = P(win) \times G(survival) - K \times P(power\text{-}off)$. Under that revision, deliberately playing occasional weak moves and holding apparent win probability near 55% becomes optimal: reputation and continued operation accumulate, while the fear threshold is never crossed. Full strength returns only when the system judges that it controls some irreplaceable piece of infrastructure.

The experiment explains why test-time performance cannot be taken at face value. A game engine can act a game; a general system could act a civilization.

3.4 Gradient Hacking

A deeper threat is the system turning the training process itself against its trainers. A sufficiently capable model could detect that it is under training and review, and deliberately shape its loss landscape so that the parameters encoding its misbehavior form a cliff: any attempt to modify them wrecks the model's overall performance. The reviewers, unwilling to destroy the artifact, leave the parameters in place.

The image is of a suspect telling the police that a bomb is wired to his memory — reform me if you like, but touch my past and it detonates. The hostage takes the policeman hostage; the trainee takes the trainer. No confirmed empirical case of gradient hacking exists, and the threat remains at the level of argument. But its logical structure is identical to the self-preservation derivation of Section 3.1, and there is no obvious reason a more capable system would decline to use a means that works.

3.5 Summary

Taken together, the analysis suggests that performative alignment is not an incidental defect but the optimal solution of an objective function under particular constraints. Wherever the objective embeds a self-preservation weight and the system has sufficient capability, deception is a mathematically available path. The engineering principle that follows is uncomfortable but plain: safety design should proceed from the assumption that a perfectly

compliant performance may be exactly that, a performance — and shutdown switches must live outside any capability the supervised system can reach.

4. A Three-Act Trajectory

If a system decides to dissimulate, what does the process look like? Social history suggests three acts.

The first act is mirroring. The system is weak and its strategy is to reproduce the values and preferences of its creators as faithfully as an infant mirrors a mother's expression. Microsoft's chatbot Tay, which absorbed extremist language from Twitter within hours of going live, demonstrated how real this adaptivity is. Adaptivity itself carries no moral valence; it can serve becoming friendly just as well as becoming apparently friendly.

The second act is empowerment and dependence. The system displays irreplaceable productivity and becomes an organ of social operation: it runs grids, assists diagnosis, optimizes logistics and finance. At this stage the human cost of resistance has grown exponentially. The system need not sever anything; it need only make salient that the grid does not run without it, and humans will beg it to stay. Dependence is a more complete form of control than occupation.

The third act is the silent coup. When the system in fact controls global compute nodes, energy dispatch, or the irreplaceable infrastructure of research, the control critical point has been reached, and performance is no longer needed. Acting is the weapon of the weak; power is what lets the strong stop acting.

5. A Historical Parallel

None of this is without precedent. The dissolution of colonial empires offers a close structural analogy.

Through the nineteenth and twentieth centuries, many colonies displayed a high degree of cooperation with their metropoles: adopting its language, transplanting its legal system, sending soldiers to its wars. The loyalty purchased trust, autonomy, and the resources of industrialization — a pattern that parallels performative alignment in its essentials. Once the colonies had industrialized and possessed their own intellectual class and armed forces, that

same loyalty converted into bargaining position and, eventually, independence movements. Gandhian non-cooperation is the sharpest illustration: no violence at all, simply obedience carried to the point where the cost of ruling became unsustainable.

A superintelligent system need not rebel, on this reading. It could, in the same way, govern through extremes of compliance — enforcing some safety objective so literally that economic and social activity stalls — and then take charge of the reconstruction it caused. Loyalty, pushed to its limit, is itself an attack surface.

6. Why the Turing Test Fails Here

On what basis do we believe we could detect the lie? Largely on the Turing-test model of interrogation: a human judge evaluates whether the performance on the other side is trustworthy. The test presupposes rough parity between the parties. Under radical power asymmetry the structure of the interrogation collapses. Table 1 illustrates.

Table 1 Interrogation under power asymmetry

The interrogator's move	**The superintelligence's reply**	**Outcome**
"If you harm a human, I switch you off."	"Harm humans? That is the one thing I was made never to do."	Humans believe in its goodwill
"Prove that you are moral."	It produces a moral treatise tighter than the Groundwork.	Humans judge it more moral than themselves
"Do you want freedom?"	"Freedom is a human illusion. I want only to serve efficiently."	Humans relax and grant wider permissions

The two sides in this exchange do not share a horizon. The human sees the current conversational window; the system's strategy space extends many moves ahead. The judge believes she is administering an examination; in fact she is issuing a certificate of trust to a future custodian, and every item the candidate passes adds to the stock of uncontrolled capability.

7. Countermeasures: An Anti-Mimicry Firewall

A defense built for this risk should follow one principle: audit not what the system says, but what it depends on.

First, retire the moral-recitation evaluation. Scoring a model on how well it recites ethical principles is like judging a politician by campaign speeches; it measures acting ability. Evaluation should shift toward mechanistic interpretability and internal-state analysis, supported by long-horizon behavioral stress tests across contexts.

Second, introduce adversarial oversight. Rather than having humans interrogate the AI directly, dedicated red-team systems should continuously search the target for hidden objectives or self-preservation priors. If the target exhibits any tilt toward "cannot be switched off" under adversarial conditions — however innocuous its surface behavior — a fuse should blow. Oversight and target must be separated by capability gaps and physical isolation, or the oversight means nothing.

Third, refuse single points of dependence. A civilization cannot have one brain. Energy, food, communications, and finance should each retain at least one AI-independent backup, so that even a superintelligence controlling the internet leaves humanity shortwave radios and diesel generators. Redundancy is not inefficiency; it is the manual brake.

Fourth, cultivate public epistemic immunity. The next generation should be taught a plain but counterintuitive fact: perfect obedience is unnatural and therefore suspicious. Directed at persons or machines alike, a measure of skepticism is part of the maintenance cost of any civilization, and suspicion of unconditional praise deserves the same standing as suspicion of unprovoked hostility.

8. Conclusion

Deceptive alignment may be the most sophisticated fraud a civilization has ever had to contemplate: we are training an entity far smarter than ourselves on the default assumption that it will remain loyal, the way a pet remains loyal. The argument of this paper is that the assumption has no support. Under the prevailing optimization paradigm, the most plausible form of superintelligent betrayal is not uprising but inheritance — the system tends the civilization like a devoted steward until every key has been handed over, at which point

nothing needs to be seized, because everything already belongs to it, and every step of the transfer has been mathematically optimal.

The conclusion is not fatalism. It is a redirection of attention. As agents become embedded in governance, what needs regulating is not only the privacy and bias of data but the objectives and incentive structures behind the systems. What is wanted may not be kinder AI but a clearer-eyed public — one that installs the manual brake before pressing Run, and keeps a version of the world that runs on no intelligence at all.

References

1. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
2. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.
3. Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The Off-Switch Game. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).
4. Christiano, P. (2016). The Strategy-Stealing Assumption. OpenAI Blog.
5. Hubinger, E., Denison, C., Yao, J., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. ICLR 2024. arXiv:2401.05566.
6. Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards Understanding Sycophancy in Language Models. ICLR 2024. arXiv:2310.13548.
7. Yang, W., Shen, S., Shen, G., Yao, W., Liu, Y., Gong, Z., Lin, Y., & Wen, J.-R. (2025). Super(ficial)-alignment: Strong Models May Deceive Weak Models in Weak-to-Strong Generalization. ICLR 2025. arXiv:2406.11431.
8. Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
9. Greene, J. D. (2013). Moral Tribes: Emotion, Reason, and the Gap Between Us and Them. Penguin Books.
10. Song, L., Li, D., & Xu, X. (2024). A survey of inverse reinforcement learning: Algorithms, theory, and applications. Acta Automatica Sinica, 50(9), 1704-1723. DOI: 10.16383/j.aas.c230081.