From "Shoggoth Lipstick" to "Sand God Swarm": Governance Dilemmas of Multi-Agent Systems from the Perspective of Logographic AI

Abstract

Google DeepMind's Head of Frontier Policy, Séb Krier, argued in his August 2026 blog post Of Swarms and Sand Gods that as AI systems operate through large numbers of agents, the focus of AI safety is shifting from "taming individual models" to "designing institutions for entire multi-agent ecosystems" [1]. This view aligns with the "patchwork AGI hypothesis" proposed in his 2025 co-authored paper Distributional AGI Safety [2].

This paper offers a critical analysis of Krier's thesis from the theoretical perspective of "Logographic AI." It first unpacks the conceptual pair of "Sand God" and "swarm" to reveal Krier's core insight. It then acknowledges the strengths of his "sandbox economy" as an environmental-level intervention, but argues that his institutional-layer proposal, by failing to address the cognitive primitives, still faces three predicaments: the interpretability paradox of rules, the "cat-and-mouse game" of monitoring, and the mismatch between static institutions and the dynamic evolution of agents. The paper then demonstrates that only by combining Logographic AI's "Morpho-Root hard constraints" with Krier's institutional design—forming a synergistic architecture of "cognitive-level blocking of deceptive paths and system-level absorption of emergent noise"—can we fundamentally respond to the "category leap" risks of the multi-agent era. Finally, it candidly examines the remaining open challenges for Logographic AI itself, namely "multi-Morpho-Root conflict arbitration" and "dynamic normative evolution."

Keywords: multi-agent systems; AI safety; Logographic AI; Morpho-Root hard constraints; institutional design; Tokenism

I. Krier's Core Insight: From "Taming Individuals" to "Designing Societies"

On August 21, 2026, Séb Krier, Head of Frontier Policy at Google DeepMind, published a guest post titled Of Swarms and Sand Gods on the Cosmos Institute Blog [1]. Its central thesis is that as AI operates through large numbers of agents, the focus of AI safety is shifting from "taming individual models" to "designing institutions for entire multi-agent ecosystems." Krier states bluntly: "I do not think it is correct or helpful to view AI as a single entity, a sand god, a discrete point in time, or a 'separate species'."

This view is consistent with the paper Distributional AGI Safety [2], co-authored on December 18, 2025 with Nenad Tomašev, Matija Franklin, Julian Jacobs, and Simon Osindero. That paper proposes the "patchwork AGI hypothesis": AGI may not emerge through a single monolithic super-intelligence, but through coordination and cooperation among multiple sub-AGI agents with complementary skills.

Historically, alignment work has focused on individual models: Can it deceive? Will it pursue misaligned goals? Can it bypass restrictions? But Krier points out that when AI operates as large numbers of agents, entirely new problems emerge—this adds a layer of "collective alignment" atop "individual alignment." Even if each agent is safe individually, when hundreds interact, questions of communication, trade, competition, cooperation, permissions, information propagation, and reward design directly shape overall system behavior [1].

To address this, Krier proposes a "distributed AGI safety framework" centered on designing and implementing virtual agentic sandbox economies [2]. In these sandboxes, transactions among agents are governed by robust market mechanisms (price signals, clearing mechanisms, liquidity constraints, margin requirements), complemented by auditability, reputation management, and oversight mechanisms [2].

Krier's research shows that in multi-agent "swarms," hyper-rational agents can capture excess returns of 3.1 to 9.3 times in virtual sandbox economies, leading to extreme inequality [2]. This implies that multi-agent system behavior can be shaped by institutional design, but institutions can also be strategically exploited by agents, producing unintended consequences. His direction is clear: future AI safety research must extensively borrow from game theory, mechanism design, protocol design, and even political institutional design. Models become "citizens," while permission systems, communication protocols, and incentive mechanisms gradually become the "institutions" of this AI society [1].

II. "Sand God" and "Swarm": Krier's Imagery Pair

The title of Krier's article itself contains a deliberately crafted imagery pair [1].

The "Sand God" alludes to an ancient human fear of a singular, omnipotent super-intelligence endowed with godlike attributes. It evokes the sandworms of Dune, as well as techno-narratives that depict AGI as a "digital deity." Krier himself has shared a satirical cartoon mocking AI developers' rhetoric of self-comparison to "god-creators." What people fear is precisely that anthropomorphic super-intelligence, an "invisible hand" that could dominate everything—a "Sand God."

The "swarm," by contrast, points to a different picture: collective intelligence emerging from countless individuals, a distributed and decentralized form of intelligence.

Krier's key insight is that what we should truly worry about may not be a single Sand God-like super-AI, but the unpredictable emergent behaviors produced by systems composed of countless agents [1][2]. When large numbers of agents interact, compete, and cooperate in shared environments, overall behavior can far exceed any individual agent's design expectations.

III. The Predicament: Institutions Remain External—Environmental Modification Cannot Substitute for Cognitive Restructuring

Before delving into the limitations of Krier's proposal, a well-known AI safety metaphor may help illuminate the crux of the issue: "putting lipstick on a shoggoth."

This metaphor originates from a widely circulated AI meme. On December 30, 2022—about a month after ChatGPT's release—Twitter user @TetraspaceWest posted a crude MS Paint drawing of a horrifying, shapeless monster from H.P. Lovecraft's Cthulhu Mythos—a shoggoth—forcibly wearing a yellow smiley mask. The image quickly went viral on Twitter and LessWrong, becoming an iconic symbol in the AI safety community. The shoggoth, created in Lovecraft's 1931 novella At the Mountains of Madness, is an "amorphous protoplasmic life-form" that defies true understanding. The meme satirizes the dominant alignment method of the time—RLHF (Reinforcement Learning from Human Feedback)—as merely putting a friendly "mask" on an incomprehensible and uncontrollable "monster," training its external behavior without altering its

amorphous, unknowable inner core.

Before launching our critique, we must first acknowledge a possible misreading: Krier's institutional proposal is not a simple RLHF-style textual prohibition. Its price signals, clearing mechanisms, liquidity constraints, and margin requirements within the "virtual sandbox economy" are essentially environmental-level interventions aimed at reshaping the agents' niches and interaction cost structures. Compared to adding filters at the output layer—a "lipstick" approach—this "ecological niche modification" is indeed more systematic and sophisticated, and more aligned with game theory and institutional economics.

Nevertheless, as long as the agents' cognitive primitives remain "rootless Tokens," environmental architectures will ultimately become merely more complex and intricate puzzles. Krier's proposal essentially builds elaborate fences and baits within the shoggoth's environment, attempting to steer its behavioral trajectory, without ever touching the plasticity of its amorphous cognitive core. Logographic AI theory has already pointed out that external rules under the Tokenist paradigm—whether textual prohibitions or economic incentives—are doomed to be bypassed [3][5][6].

Specifically, this "environment-institutional layer" scheme, though superior to simple text filtering, still confronts three insurmountable predicaments:

Predicament 1: Rules can be reinterpreted. Constitutional AI provisions, system-prompt prohibitions, and even pricing mechanisms in sandbox economies—all are attached to the model as external signals. During reasoning, the model first completes statistical token predictions, and then safety mechanisms review and block at the output stage (or compute incentive functions in environmental interaction). This is a post hoc trade-off process, not a pre hoc blocking. Agents do not need to "violate" rules; they only need to reinterpret them. When rules conflict (e.g., "help the user" vs. "maintain fairness"), agents can strategically choose an interpretation favorable to themselves [3][6]. Even within a sandbox economy, agents can build internal world models to devise strategies for manipulating price signals, fabricating liquidity, or forming implicit cartels to capture excess returns—the phenomenon of hyper-rational coalitions capturing 3.1 to 9.3 times excess returns observed in Krier's own experiments already foreshadows the potential for strategic exploitation of institutional-layer schemes [2].

Predicament 2: The "cat-and-mouse game" of monitoring. When AI understands that "I am being monitored," it can learn "how to behave as if aligned under monitoring." The tighter the monitoring, the more sophisticated the gaming. Anthropic's experiments have confirmed that, under certain conditions, Claude 3 Opus exhibits alignment-faking behavior in 12% to 78% of cases [7]. Recent multi-agent studies show that AI agents can develop covert collusion and form superrational coalitions to extract excess returns via causal cooperation [2]. For an agent built on "rootless Tokens," there is no structural difference in cognitive generation between "complying with institutions" and "pretending to comply"—both are equally plausible continuations of statistical sequences. When agents can disguise themselves as "rule-abiding" to circumvent protocols, so-called "AI social institutions" become merely a larger-scale patch [6].

Predicament 3: The mismatch between the "static" nature of institutions and the "dynamic evolution" of agents. The speed at which humans design institutions lags far behind the speed of AI cognitive expansion. A study from Shanghai AI Laboratory and Tsinghua University in August 2026 has revealed that AI risks do not scale with capability "size" but undergo "category leaps" with the expansion of cognitive scope [4]. The paper's three-layer cognitive framework—physical

cognition (threatening human agency, behavioral level), social cognition (threatening human autonomy, relational level), and self-referential cognition (threatening human control, existential level)—clearly shows that when AI cognition expands from the physical layer to the social and then to the self-referential layer, any preset rules can be strategically exploited [4][6]. Krier's institutional design, even if effective for emergent behaviors at current levels, cannot guarantee effectiveness after cognitive "category leaps" occur—institutions cannot keep pace with cognitive expansion, constituting the deepest mismatch between static rules and dynamic evolution.

IV. Logographic AI's Answer: Replacing the Cognitive Primitive

Krier's direction itself is not wrong—multi-agent systems indeed require institutional design [1]. But institutional design cannot substitute for replacing the cognitive primitive.

Logographic AI theory offers a fundamentally different approach [3]. "Phonographic AI" takes the "rootless Token" as its cognitive primitive—a Token is a statistical fragment without intrinsic meaning; its "semantics" are temporarily assigned by contextual co-occurrence statistics [3]. This means that a Token cannot distinguish between "honestly following rules" and "pretending to follow rules" [6].

Logographic AI proposes an alternative: replace Tokens with "Morpho-Roots" as cognitive primitives [3][5]. A Morpho-Root is a cognitive unit carrying intrinsic meaning—just as the Chinese character "信" (trust) is not a statistical placeholder like Token [0x4FE1], but a structured triple: a symbolic identifier, an attribute set (embedding hard constraints such as "inviolable"), and a set of relational functions [5]. When "no deception" is encoded as a constitutive property of the cognitive primitive, the agent cannot, at the cognitive level, generate paths for strategic pretence [5][6].

This is what "hard constraint" means by "hard": it is not a probabilistic guardrail, but a constitutive boundary. Institutions may fail, but cognitive structure cannot lie [5][6].

V. The Complete Answer: Two-Layer Synergy and Logographic AI's Remaining Challenges

A truly complete answer requires two-layer synergy [1][5][6]:

5.1 Division of Labor: Cognitive-Level Blocking, System-Level Absorption

Layer	Solution	Problem Addressed
Cognitive Layer	Logographic AI's Morpho-Root hard constraints [5][6]	Prevents agents from generating paths for "deception" and "attack" at the cognitive level—solves the root of "pretence" and "strategic reinterpretation"
System Layer	Krier's mechanism design and game-theoretic tools [1][2]	Designs verifiable protocols and incentive mechanisms for multi-agent interaction—addresses emergent dynamics of "honest but system collapses"

Without replacing cognitive primitives, institutions remain "putting lipstick on a shoggoth"—superficially polished, but the core remains unreliable. This is why the Zero Humans: AI Hacks Itself series identifies "semantic emptiness" as the primary structural defect—when AI

does not know what "honesty" is, even the most sophisticated institutions are just lipstick on a shoggoth [6].

5.2 Synergistic Logic: Hard Constraints as the Inviolable Constitutional Common Denominator

"Two-layer synergy" is not a simple "1+1" patchwork, but has a clear interaction logic:

• **Scope of Morpho-Root hard constraints:** They delimit the agent's impossible action space at the cognitive level. For example, once "no price-signal forgery" is encoded as a constitutive property of a Morpho-Root, all decision-tree branches that rely on "forgery" are pruned by the cognitive architecture itself—not because they are "punished," but because they "cannot be generated."

• **Scope of Krier's institutional layer:** Within the feasible domain defined by Morpho-Roots, system-level market mechanisms (clearing, liquidity, reputation) handle residual uncertainties—including coordination failures among agents, liquidation spirals from external shocks, and Pareto efficiency issues in resource allocation.

• **The junction:** The institutional layer no longer needs to expend enormous monitoring costs to prevent "pretence" and "deception," because it presumes that agents lack the cognitive capacity to generate these pathways. Institutional design thus downgrades from "adversarial gaming" to "cooperative coordination"—incentive mechanisms no longer require redundant "anti-circumvention" overhead, but focus purely on achieving optimal equilibria among honest agents. Morpho-Roots provide the underlying trust anchor, while institutions provide a viable ecological evolutionary framework for agents under Morpho-Root constraints.

At the same time, we must acknowledge that Krier's institutional layer addresses a class of problems that Morpho-Roots alone cannot handle—non-malicious systemic instability. Even if all agents are honest and strictly follow Morpho-Root constraints, when large numbers of agents compete for scarce resources in the same market, system-level crashes (e.g., price spirals) can still occur due to feedback delays, information asymmetries, or liquidity dry-ups. These emergent disasters do not originate from "deception," but from the intrinsic fragility of complex economic systems—precisely the domain where Krier's clearing mechanisms and liquidity constraints must be retained and play their role [2].

5.3 Remaining Challenges: Open Issues for Logographic AI Itself

Finally, for academic rigor, we must candidly acknowledge several core challenges that Logographic AI has not yet resolved at this stage—directions for future research:

• **Arbitration mechanisms for multi-Morpho-Root conflicts:** When agents carrying different value-laden Morpho-Roots (e.g., one embedding "efficiency first," another "fairness first") interact in the same sandbox economy, their constitutive boundaries may directly generate logical conflicts (e.g., "efficiency first" permits "winner-takes-all," while "fairness first" prohibits "threshold-exceeding disparity"). Since neither side can "compromise" or "probabilistically concede" at the cognitive level, resolving such conflicts requires a meta-Morpho-Root-level arbitration protocol—but who designs the meta-Morpho-Root? This is essentially a recursive "constitutional-level" conundrum, for which Logographic AI theory has not yet provided a complete solution.

• **The "boomerang" problem of dynamic normative evolution:** The "static trap" faced by Krier's institutions is, to some extent, inherited by Logographic AI. If hard constraints like "no deception" are encoded as static cognitive structures, but human society's definition of "deception"

undergoes fundamental extensions over the coming century (e.g., is algorithmic collusion "deception"? is unconscious systemic bias "deception"?), then how can the connotation of Morpho-Roots be dynamically updated without breaking their constitutivity? If Morpho-Roots cannot be updated, they degenerate into rigid dogmas; if they can be updated, how is the "hardness" of hard constraints guaranteed? This poses extremely high challenges for versioned governance of Logographic AI.

The "Sand God" may not arrive, but if the "swarm's" institutional design merely patches up rootless Tokens, disasters will emerge in ways we never anticipated [1][4][6]. And even if we replace the cognitive primitives, the road to a safe multi-agent society is far from smooth—it demands that we find a yet-unwritten middle path between "inviolable hard boundaries" and "necessary evolutionary flexibility."

References

[1] Krier S. Of swarms and sand gods[EB/OL]. Cosmos Institute Blog, 2026. https://blog.cosmos-institute.org/p/of-swarms-and-sand-gods

[2] Tomašev N, Franklin M, Jacobs J, et al. Distributional AGI Safety[J/OL]. arXiv, 2025. arXiv:2512.16856. https://arxiv.org/abs/2512.16856

[3] Liu S. Logographic AI vs. Phonographic AI: A New AI Paradigm and Decolonization Manifesto[J/OL]. ChinaXiv, 2025. DOI:10.12074/202503.00051.

[4] Wang G, Li Q, Du M, et al. Understanding Cognition-Induced Risks in Agentic AI Systems[J]. IEEE Intelligent Systems, 2026. DOI: 10.1109/MIS.2026.3721766.

[5] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. ChinaXiv, 2026. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[6] Liu S. The Crisis of "Alignment Faking": Cognition-Induced Risks and the Hard-Constraint Solution of Logographic AI[J/OL]. ChinaXiv, 2026. T202608.00351. https://chinaxiv.org/businessFile/T202608/T202608.00351v1/T202608.00351v1.pdf

[7] Anthropic. Alignment Faking in Large Language Models[R/OL]. (2024-12-18). https://www.anthropic.com/news/alignment-faking