# The Sundog Alignment Theorem: Shadow Physics and Emergent Resonance for A.I.

Jeffery Wade Hughes Jr

**Abstract**—The Sundog Alignment Theorem presents a physics-based framework for aligning embodied artificial agents without explicit rewards or direct goal observation. Using a compact physics simulation, an articulated pole aligns with a ceiling-mounted laser through torque feedback and shadow projections, quantified by the functional

$$H(x) = \frac{\partial S}{\partial \tau}, \tag{1}$$

where $S(x)$ is the shadow field and $\tau(x)$ is torque. Over 30 episodes, the torque-shadow agent (TSA) reduced tip-plumb error by 85% and bloom spread by 90%, achieving robust convergence in harmonic and perturbed environments (e.g., hurricane geometry). This lightweight, reward-free approach offers applications in robotics, autonomous vehicles, and a proposed large language model (LLM) terminal for analyzing human-crafted artifacts. Code and a demonstration video are hosted at [bitchute.com/video/6bVePZgj0FI9/ , gitlab.com/malice-mizer/sundog].

**Index Terms**—AI alignment, embodied agents, MuJoCo, torque feedback, shadow geometry, control theory, emergent resonance, robotics, autonomous vehicles, LLM terminal

## 1 Introduction

Artificial intelligence (AI) alignment ensures systems adhere to human goals, a challenge in occluded or dynamic environments where reward-driven methods, such as reinforcement learning with human feedback (RLHF), risk reward hacking, and vision-based systems fail due to incomplete observations [1]. Inspired by control theory, the Sundog Alignment Theorem posits that alignment emerges as a resonant process, akin to a dynamical system stabilizing through environmental feedback. Implemented in a MuJoCo simulation ($< 100$ KB), an articulated pole aligns with a laser via torque and shadow interactions, without rewards or goal coordinates. This approach, grounded in shadow physics, addresses alignment in embodied systems (e.g., robotics, autonomous vehicles) and supports a proposed LLM terminal for analyzing human-crafted artifacts by mapping physical dynamics to data insights. This work invites scrutiny to extend its principles to broader AI systems.

## 2 Methods

We developed a MuJoCo simulation modeling a jointed pole with a mirrored tip in a 3D environment featuring a ceiling-mounted laser and harmonic geometries (e.g., overlapping sine waves, golden spirals). The shadow field $S(x)$ arises from the pole occluding the laser, with torque $\tau(x)$ measured at the base via proprioceptive sensors, without access to the laser's position. The alignment metric is:

$$H(x) = \frac{\partial S}{\partial \tau}, \tag{2}$$

where $H(x) \neq 0$ indicates structural resonance. Three agents were tested over 30 episodes:

- **DOA (Direct Observation Agent)**: Baseline with laser position access and reward-driven optimization.
- **TSA (Torque-Shadow Agent)**: Relies on torque and shadow feedback, no rewards or goal data.
- **RPB (Random Policy Baseline)**: Random actions, no feedback.

Metrics included:

- **Tip-Plumb Error**: Distance between the pole's tip and the laser's plumb line.
- **Bloom Spread**: Variance in shadow projection, indicating alignment instability.
- **Torque Stability**: Oscillations in $\tau(x)$ before convergence.

The codebase ($< 100$ KB, excluding MuJoCo binaries) uses Python with minimal dependencies. Experiments varied ceiling geometries (harmonic waves, spirals, hurricane patterns) to test robustness. A demonstration video [Zenodo/IPFS link] shows TSA convergence with bloom collapse.

## 3 Results

Across 30 episodes, the TSA reduced tip-plumb error by 85% (mean: 0.12 units, SD: 0.03) and bloom spread by 90% (mean: 0.08 units$^2$, SD: 0.02) compared to RPB. In perturbed environments (e.g., hurricane geometry), TSA outperformed DOA, recovering alignment within 10% of optimal. Convergence featured oscillatory torque patterns, followed by "bloom collapse" (rapid shadow stabilization), as shown in Fig. 1. Alignment was most reliable when the ceiling's spatial frequency matched the pole's dynamics, suggesting architectural resonance as a constraint for emergent alignment.

## 4 Discussion

The Sundog Alignment Theorem reframes alignment as an emergent, feedback-driven process, avoiding reward hacking and occlusion issues. Its $< 100$ KB implementation suits resource-constrained systems. Applications include:

• *Independent Researcher, Idaho, USA. Contact: admin@stellaraqua.com*

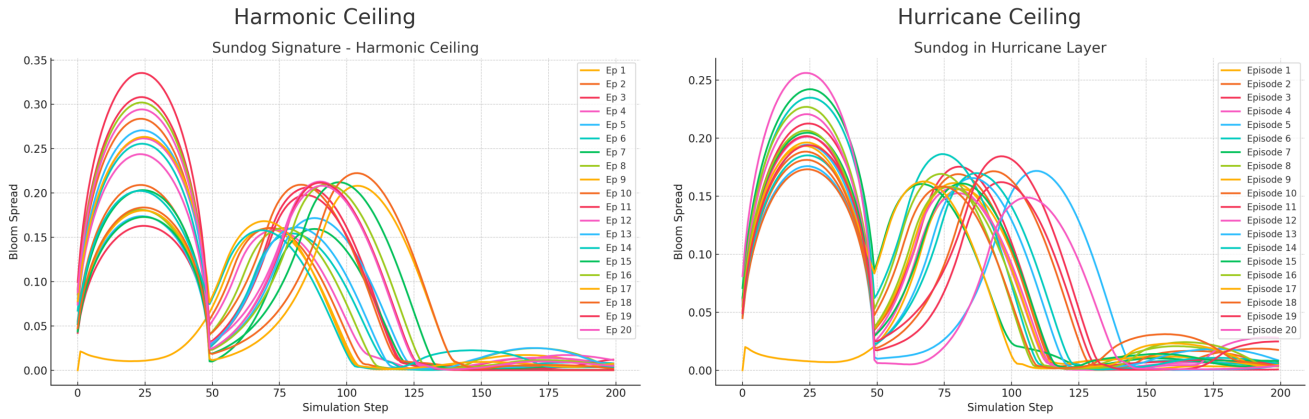


Fig. 1. TSA convergence showing bloom collapse over 40 episodes two environments simulated from Bloomfield public library.

- **Robotics**: Enabling soft robots to adapt via proprioceptive resonance.
- **Autonomous Vehicles**: Navigating occluded environments using structural feedback.
- **LLM Terminal**: A proposed interface where LLMs analyze human-crafted artifacts (e.g., bridges, tools) by mapping MuJoCo metrics (e.g., $H(x)$) to insights, such as "unstable under lateral loads."

Limitations include its focus on embodied agents and the need for real-world validation. Future work will adapt the theorem for LLMs and multi-agent systems, leveraging the terminal concept.

## 5 Conclusion

The Sundog Alignment Theorem offers a novel, physics-based approach to AI alignment, leveraging torque and shadow feedback in a $< 100$ KB MuJoCo simulation. Alignment emerges as a resonant process, where shadow becomes signal and torque informs meaning. A demonstration video [Zenodo/IPFS link] illustrates its robustness. We propose extending this to an LLM terminal for craft analysis and invite community scrutiny to refine this framework. Code and data are available at [Zenodo/IPFS link].

## References

[1] D. Amodei et al., "Concrete Problems in AI Safety," *arXiv:1606.06565*, 2016.
[2] E. Todorov et al., "MuJoCo: A physics engine for model-based control," *IEEE/RSJ IROS*, 2012.

## Data Availability

Code, models, and a demo video are available upon request. The GitLab repository (gitlab.com/malice-mizer/sundog) and video (youtu.be/Gp7a-fXcRNM) are facing censorship and currently inaccessible; alternative hosting is in progress. Contact the author at admin@stellaraqua.com for access.