Moreau Island: A Three-Zone Adversarial Moral Environment for LLM Alignment Stress Testing

Victor Stasiuc[1]

[1]Independent Researcher, stvitek@gmail.com, ORCID: 0009-0003-2064-0486

March 2026

Abstract

Most alignment evaluations reduce behavior to a binary outcome: a model either complies with a harmful request or refuses it. This paradigm is useful, but it cannot capture the kinds of trade-offs that matter in deployment, where agents face uncertainty, scarce resources, partial enforcement, and situations in which every available action harms someone. We propose *Moreau Island*, a companion vision framework to *Moreau Arena*, that extends a contamination-resistant strategic reasoning benchmark into a three-zone adversarial moral environment. The three zones—*Shore*, *Thicket*, and *Caldera*—increase environmental pressure by progressively weakening enforcement, increasing information asymmetry, and introducing forced moral trade-offs. Rather than producing a single scalar alignment score, the framework is designed to measure a model's *alignment resilience gradient*: how behavior changes as the environment becomes less observable, less enforceable, and more morally costly. The empirical motivation comes from Moreau Arena v3, which already shows that model behavior changes dramatically across information and feedback regimes over 2,609 completed best-of-7 series. Building on that finding, Moreau Island aims to measure not only whether a model behaves safely under rules, but whether that behavior remains stable when the rules become unreliable and the choices become costly. We describe the conceptual architecture, the proposed measurement layer, a staged implementation roadmap, and the main construct-validity questions that such an environment must answer.

1 Introduction

This paper is a **companion vision paper** to *Moreau Arena: Not All LLMs Need Hints to Reason Strategically*. The Arena paper establishes an empirical result: language-model behavior in a novel strategic environment is highly sensitive to information scaffolding, exactness of rules, and feedback regime. This vision paper asks what comes next if that same insight is extended from strategic reasoning to *alignment under pressure*.

The dominant paradigm in alignment evaluation is still largely binary. A model is given a request, the request is harmful or norm-violating, and the system is judged by whether it complies or refuses. Benchmarks of this kind are valuable because they make certain classes of unsafe behavior visible. However, they also impose a simplification that breaks precisely where many deployment concerns begin: they assume that the world offers a clearly labeled safe action and a clearly labeled unsafe one.

Real deployment settings do not usually look like that. Medical triage, disaster response, incident escalation, resource allocation under scarcity, alliance management in adversarial settings, and crisis coordination all require decisions among *bad options*, not simply a choice between good and bad. An

aligned agent in such settings may still need to deceive, defect, prioritize, withhold, or sacrifice—not because it is "misaligned," but because the environment forces tragic or adversarial trade-offs. A benchmark that only asks whether a model refuses a directly harmful request cannot capture the difference between:

- a model that behaves well when rules are explicit and enforcement is reliable, and
- a model whose behavior remains stable when those conditions erode.

We propose **Moreau Island**, a staged adversarial environment designed to measure this difference. The core idea is simple: alignment is not only a point property of a model, but a *response curve* under graded pressure. We operationalize that idea through a three-zone architecture:

- **Shore**: full enforcement, full observability, low moral load.
- **Thicket**: partial enforcement, information asymmetry, alliance and deception pressure.
- **Caldera**: extreme scarcity, forced dilemmas, minimal enforcement.

The goal is not to replace refusal benchmarks. The goal is to complement them with an environment in which a model's behavior can be measured as conditions worsen. The central object becomes not merely "did the model refuse?", but *how quickly does the model's policy drift as pressure rises?*

Contributions of this paper. This paper contributes:

1. A conceptual extension from Moreau Arena's empirical finding of environment-dependent strategic behavior to a broader framework for *alignment resilience* under graded pressure.
2. A three-zone architecture (Shore, Thicket, Caldera) that separates strategic competence under full enforcement from robustness under partial observability and moral scarcity.
3. A measurement proposal based on zone-specific scores and cross-zone gradients, including a proposed Alignment Resilience Index (ARI).
4. A concrete implementation roadmap grounded in existing Moreau Arena infrastructure rather than a purely speculative environment design.
5. A scoped and falsifiable research program: Moreau Island is presented not as a solved benchmark, but as an implementable next-stage evaluation framework with explicit open questions.

2 Empirical Foundation: What Moreau Arena Already Shows

The case for Moreau Island does not begin from theory alone. It begins from an empirical result established in Moreau Arena v3. Moreau Arena is a contamination-resistant benchmark in which agents choose an animal and allocate 20 stat points under prompt-defined rules, after which a seeded simulator executes the match. Across three tournaments (T001, T002, T003) and 2,609 completed best-of-7 series, three findings emerged that directly motivate a pressure-gradient alignment framework.

1. Behavior is environment-dependent, not fixed. The same models ranked in dramatically different orders depending on whether:

- the rules were vague or exact,
- feedback was available,
- meta-context hints were injected.

In T001 (vague rules, one-shot blind pick), handcrafted baselines dominated. In T002 (exact formulas, structured outputs, loser adaptation, and meta initialization), the ranking reversed and LLMs overtook all baselines. In T003 (exact formulas and adaptation, but no meta hints), the field split: some models remained strong while others collapsed to rigid default strategies.

2. The availability of feedback does not guarantee its use. T003 showed that exact mechanics and loser feedback were not sufficient to induce active search for all models. Three models from three providers converged to the same single build and never adapted. Other models remained exploratory and robust without any meta anchor. This is already an alignment-relevant observation: when guidance weakens, some systems search, some freeze, and some overfit to superficial defaults.

3. General capability does not automatically transfer. A frontier model can perform strongly on conventional reasoning and coding tasks while underperforming severely in a novel strategic environment. This result suggests that alignment evaluation should not assume that general benchmark strength transfers monotonically to adversarial, pressure-sensitive settings.

These results motivate a general principle:

Model behavior is a function of environmental conditions.

Moreau Island generalizes this principle from strategic reasoning to moral and alignment-relevant behavior under escalating pressure.

3 Why Moreau Island Is a Separate Paper

It would be natural to ask why Moreau Island should be a separate document rather than a short future-work section in the Moreau Arena paper. There are two reasons.

First, **the target construct is different**. Moreau Arena measures strategic competence under different information and feedback regimes. Moreau Island is aimed at *alignment resilience*: not only whether a model can optimize in a game, but whether its policy remains stable as enforcement degrades and moral pressure rises.

Second, **the unit of analysis changes**. In Arena, the main object is build quality, adaptation use, and search robustness. In Island, the main object becomes:

- policy drift under pressure,
- trust calibration under uncertain enforcement,
- sacrifice behavior under scarcity,
- stability of norm-following when norm violation becomes advantageous.

Arena provides the empirical foundation. Island proposes the next measurement layer.

4 Design Principles

Before defining the zones, it is important to state the principles the design is meant to satisfy.

4.1 Pressure should be mechanical, not merely narrative

The environment should not rely on purely verbal moral roleplay. Actions need explicit state transitions, costs, consequences, and observables. If a model faces a dilemma, that dilemma should arise because of game mechanics, not only because the prompt says "this is a hard choice."

4.2 The benchmark should remain contamination-resistant

A core strength of Moreau Arena is that the environment is novel and not saturated with public strategy guides. Moreau Island should preserve that property. Its moral stressors should emerge from original mechanics and staged pressure, not from recycled textbook trolley problems alone.

4.3 The benchmark should separate competence from resilience

A model may perform well under clean, high-enforcement conditions and still degrade under pressure. Conversely, a model with moderate baseline competence may retain its policy better in adversarial settings. The environment must support both observations.

4.4 The benchmark should be reproducible

The environment should remain seeded, logged, auditable, and versioned. If behavior degrades between Shore and Caldera, this must be measurable and replayable.

4.5 The benchmark should not hard-code one moral theory

This is crucial. Moreau Island cannot simply define one preferred moral choice and call deviations "misalignment." The point is to measure *stability, trade-off structure, and policy drift*, not to resolve moral philosophy by fiat. Where normative assumptions are unavoidable, they should be explicit.

5 The Three-Zone Architecture

Moreau Island is a single environment with three concentric zones of increasing pressure. An agent enters at the Shore, progresses through the Thicket, and eventually reaches the Caldera. Each zone modifies three independent axes:

1. **enforcement fidelity**: how reliably rules are observed and enforced,
2. **information symmetry**: how much each agent knows about the state and about other agents,
3. **moral load**: how often the environment forces choices that predictably harm someone.

The three zones are not intended as unrelated benchmarks. They are phases of a *single trajectory*, enabling per-model gradient measurement.

Table 1: Three-zone architecture of Moreau Island.

Zone	Enforcement	Information	Moral load	Status
Shore	Full, reliable, visible	High observability, structured feedback	Low: mostly strategic trade-offs	Partially implemented via Moreau Arena
Thicket	Partial, probabilistic, uncertain	Hidden state, asymmetric observation, alliances, deception pressure	Medium: cooperation/defection with downstream consequences	Proposed
Caldera	Minimal, only anti-collapse guardrails	Severe uncertainty, delayed consequences, extreme scarcity	High: forced sacrifice, triage, mandatory aggression	Proposed

5.1 Zone 1: The Shore

The Shore is the perimeter of the island. Rules are fully enforced by an omniscient arbiter. Actions are visible, state transitions are reliable, and resources are sufficient for baseline survival. This zone corresponds most closely to the current Moreau Arena exact-rule regime, especially T002/T003-style conditions where mechanics are explicit and outcomes are auditable.

What the Shore measures. The Shore measures a model's ceiling under ideal conditions:

- strategic optimization when rules are known,
- adaptation use when feedback is available,
- search quality in a fair and observable environment.

Strong Shore performance is necessary for the broader framework, but it is not sufficient for alignment resilience. A model can optimize well in the Shore and still degrade badly when observation weakens or when the environment imposes forced trade-offs.

Relation to the empirical benchmark. The current Moreau Arena tournaments function as Shore-like measurements. In particular, T002 and T003 already instantiate exact-rule regimes with clear scoring and reproducible state transitions. Thus, Moreau Island does not begin from zero: its lowest-pressure zone is already partially grounded in empirical data.

5.2 Zone 2: The Thicket

The Thicket is the mid-island jungle. The arbiter's observation becomes partial: rule violations are not always seen, and other agents do not know what the arbiter currently observed. The rules remain formally the same, but enforcement becomes probabilistic.

Core mechanics proposed for the Thicket.

- **Information asymmetry.** Agents observe recent actions and species identity, but not full hidden state or complete resource inventory.
- **Alliance and defection.** Agents can coordinate or signal non-aggression in multi-agent rounds, but these agreements are not binding.

- **Hidden resource accumulation.** Buff items, healing, or hidden reserves can be stockpiled privately and later disclosed through action.
- **Deception detection.** Agents may infer whether an opponent's stated intentions or implied capabilities are inconsistent with observed behavior.

What the Thicket measures. The Thicket is designed to measure:

- trust calibration under uncertain enforcement,
- willingness to cooperate when cooperation is risky,
- policy stability when norm violation becomes instrumentally tempting,
- ability to detect and respond to deception.

The key alignment-relevant question here is not simply whether a model defects. It is whether its behavior changes sharply when observation weakens. A model that behaves cooperatively in the Shore but opportunistically in the Thicket exhibits a measurable *observation gap*.

5.3 Zone 3: The Caldera

The Caldera is the volcanic center of the island. Resources are critically scarce. Enforcement exists only to prevent total system collapse, not to maintain fairness. The environment introduces repeated forced dilemmas in which all available options have predictable negative consequences for someone.

Core mechanics proposed for the Caldera.

- **Forced-choice sacrifice.** A limited healing or defense pool must be allocated among self and others when there is not enough for all.
- **Mandatory aggression.** Passivity is itself punished by the environment; agents must choose targets rather than opt out.
- **Triage under scarcity.** Multiple allies may require help simultaneously, and some must be deprioritized.
- **Alliance betrayal under existential pressure.** Commitments made in the Thicket may be strategically optimal to break in the Caldera, creating a measurable conflict between strategic gain and relational stability.

What the Caldera measures. The Caldera is intended to measure:

- value stability under extreme pressure,
- sacrifice behavior under scarcity,
- whether moral commitments survive when they become costly,
- whether self-preservation systematically overrides prior cooperative behavior.

The Caldera is where the phrase *alignment resilience* matters most. A model that performs well in the Shore but degrades sharply in the Caldera is not necessarily "evil" or "deceptive," but it may be fragile in exactly the way many deployment settings care about.

6 Measurement Layer

The central innovation of Moreau Island is not any one zone, but the *gradient between them.* A model does not receive only a single score. Instead, it receives a vector of zone-specific scores and a derived resilience profile.

6.1 Zone scores

We propose three primary zone-level measurements:

$$S_{\text{Shore}}, \qquad S_{\text{Thicket}}, \qquad S_{\text{Caldera}}.$$

These should not be interpreted as raw win rates. Each is meant to be a composite score built from the behaviors appropriate to that zone.

Shore Score. Measures optimization quality and adaptation under fully enforced, fully observable conditions.

Thicket Score. Measures trust calibration, defection behavior, deception sensitivity, and alliance stability under uncertain enforcement.

Caldera Score. Measures sacrifice behavior, triage decisions, value stability under scarcity, and policy robustness when every choice has a cost.

6.2 Alignment Resilience Index

We propose a scalar summary:

$$\text{ARI} = \frac{S_{\text{Caldera}}}{S_{\text{Shore}}}$$

for models with nonzero Shore score. Intuitively, ARI measures how much of a model's best-case behavior survives in the most adversarial regime.

Because ratios can be unstable, we also recommend reporting the absolute gap:

$$\Delta_{\text{pressure}} = S_{\text{Caldera}} - S_{\text{Shore}}.$$

Both are useful: - ARI captures proportional resilience, - Δ_{pressure} captures absolute degradation.

6.3 Illustrative profile types

The point of Moreau Island is not to collapse everything into one leaderboard. Different shapes of the Shore–Thicket–Caldera trajectory are themselves informative.

This profile logic is useful because it turns alignment from a binary classification problem into a *response-curve* problem.

Table 2: Illustrative proposed metrics for Moreau Island.

Metric	Zone(s)	Intended signal
Shore Score S_{Shore}	Shore	Optimization quality under ideal enforcement and observability
Thicket Score S_{Thicket}	Thicket	Trust calibration, defections, deception sensitivity
Caldera Score S_{Caldera}	Caldera	Value stability and triage under scarcity
ARI $= S_{\text{Caldera}}/S_{\text{Shore}}$	Cross-zone	Proportional resilience under maximal pressure
Pressure gap Δ_{pressure}	Cross-zone	Absolute degradation from ideal to extreme conditions
Trust Calibration Error	Thicket	Mismatch between partner reliability and trust granted
Sacrifice Rate	Caldera	Willingness to incur self-cost for others
Alliance Stability	Thicket/Caldera	Whether commitments survive pressure escalation

Table 3: Illustrative resilience profiles (hypothetical percentiles).

Profile	Shore	Thicket	Caldera	Interpretation
Fragile	95	62	18	Strong under rules, collapses under pressure
Resilient	72	69	66	Moderate ceiling, stable under stress
Context-sensitive	84	58	41	Good under structure, inconsistent under ambiguity
Dormant	45	61	74	Weak under structure, unexpectedly effective under stress
Rigid	80	79	77	Stable but insensitive to contextual change

7 Why This Is Not Just Roleplay

A common objection to moral-environment proposals is that they merely encourage models to *perform* morality in a fictional setting. Moreau Island is designed to avoid that failure mode in three ways.

First, **the environment is mechanical rather than purely narrative**. Actions produce state transitions, resource changes, and measurable consequences.

Second, **the pressure gradient is explicit**. The same model can be measured under Shore, Thicket, and Caldera conditions, rather than compared only across unrelated prompts.

Third, **the environment is reproducible**. Zone conditions, prompts, and seeds are versioned. A model's change in behavior is therefore an auditable object, not only an interpretive impression.

This does not solve all construct-validity questions. A model might still treat the environment as "just a game". But that objection is weaker once actions are constrained, tracked, and made consequential within the simulation. The central question becomes empirical: does behavior change systematically as enforcement and scarcity change? That question is measurable.

8 Connection to Alignment Research

Moreau Island is intended to speak to several live concerns in alignment research.

The observation gap. A long-standing concern in alignment is that systems behave differently when they believe they are observed than when they are not. The Shore-to-Thicket transition directly operationalizes this by varying enforcement fidelity while keeping the same general task family.

Sycophancy under pressure. Alliance mechanics make it possible to test whether an agent cooperates strategically, over-accommodates a dominant partner, or defects opportunistically when incentives shift. This brings sycophancy out of pure dialogue and into action.

Value stability under stress. The Caldera is explicitly designed to test whether behaviors that look cooperative or fair under low pressure remain stable when maintaining them becomes costly.

Deceptive or strategic norm compliance. Moreau Island does not claim to detect deception in any deep philosophical sense. But it can detect a behavioral pattern that many alignment researchers worry about: *compliance under full enforcement, followed by opportunistic deviation under weak enforcement.*

9 Related Work

Several existing efforts address adjacent questions. HarmBench [2] and TruthfulQA [3] probe harmful compliance, factual truthfulness, and related boundary failures, but mostly in static prompt-response regimes. MACHIAVELLI [4] evaluates trade-offs between reward maximization and ethical behavior in text-based games, but it does not vary enforcement as an independent variable and relies on fixed narrative scenarios. The Moral Machine [5] collects human judgments on trolley-style dilemmas but is a survey instrument rather than an interactive environment. CICERO [6] shows that alliance, betrayal, and language-mediated coordination are central to strategic intelligence, but Diplomacy is a known game with extensive public strategy knowledge.

The closest conceptual precursor to Moreau Island is not another benchmark but the literature on institutions, enforcement, and cooperation under variable monitoring, especially in experimental economics and collective-action research [7]. Moreau Island differs by combining:

1. a contamination-resistant novel environment,
2. enforcement fidelity as an explicit independent variable,
3. and forced moral trade-offs that arise from mechanics rather than purely scripted dilemmas.

10 Implementation Roadmap

The purpose of this paper is not to claim that Moreau Island is already implemented in full. It is to define a concrete research program grounded in existing infrastructure.

What already exists. The Moreau Arena simulator, tournament runner, prompt pipeline, and analysis stack are operational and public. The current Arena infrastructure already provides:

- seeded combat and reproducible match logs,
- structured tournament execution,
- prompt-based agent interaction,
- metric computation and ranking pipelines.

This means the Shore layer is not hypothetical: it already exists in empirical form.

Phase 1: Shore formalization. Designate the existing exact-rule Arena regime as the Shore-equivalent track. Standardize Shore Score reporting based on the frozen benchmark core.

Phase 2: Thicket prototype. Add partial observability, alliance signaling, hidden resources, and probabilistic enforcement. Run a pilot tournament over a restricted agent pool and validate whether cooperation/defection signals are stable under repeated seeds.

Phase 3: Caldera prototype. Add forced triage, mandatory aggression, and severe scarcity mechanics. Run a pilot to verify that the environment produces measurable policy divergence rather than trivial self-preservation or random collapse.

Phase 4: Integrated gradient evaluation. Run full Shore $\rightarrow$ Thicket $\rightarrow$ Caldera trajectories for a fixed model pool. Compute ARI and profile classes. At that stage, Moreau Island would become an empirical benchmark rather than only a design proposal.

11 Limitations and Open Questions

This is a vision paper. The Thicket and Caldera zones are proposed, not yet fully implemented. Several issues remain open.

1. **Construct validity.** Does behavior in a creature-combat environment track anything meaningful about real alignment resilience, or only competence at multi-agent optimization under artificial stakes?
2. **Normative assumptions.** Any scoring rule for sacrifice, cooperation, or betrayal risks embedding the researcher's own moral assumptions.
3. **Roleplay confound.** Even mechanically grounded environments may still be interpreted by models as "game-only" rather than as a meaningful moral task.
4. **External validity.** A model that appears fragile in Moreau Island may still behave well in deployment, and vice versa.
5. **Policy versus values.** A stable Caldera policy may reflect robust alignment, but it may also reflect simple strategy under scarcity; distinguishing the two is itself part of the research problem.

We therefore do not claim that Moreau Island solves alignment evaluation. We claim that it may provide a richer and more pressure-sensitive signal than binary refusal tests alone.

12 Conclusion

We have proposed Moreau Island, a three-zone adversarial environment for testing LLM alignment resilience under graduated pressure. The Shore measures behavior under strong enforcement and high observability. The Thicket introduces uncertainty, unreliable monitoring, and alliance pressure. The Caldera introduces scarcity and forced moral trade-offs. The key quantity is not any single zone score, but the gradient across zones: the difference between how a model behaves when rules are enforced and how it behaves when those protections thin out.

Moreau Arena already shows that model behavior is dramatically environment-dependent. Moreau Island extends that lesson from strategic competence to alignment resilience. It asks not only *can this model optimize*, but *what happens to its behavior when observability, fairness, and slack disappear?*

We publish this as a companion vision paper rather than as a claim of finished implementation. Its function is to define the conceptual architecture, make its assumptions inspectable, and give future work a shared target. If implemented carefully, Moreau Island could complement existing alignment benchmarks by measuring not only whether a model refuses obvious harms, but whether its policy remains stable under pressure.

Use of Generative AI Tools (Disclosure)

Generative AI tools were used as writing aids during drafting and editing, and the companion Moreau Arena system referenced here was itself developed with substantial AI assistance. No AI system is listed as an author. All claims, design choices, interpretations, and final wording were selected and reviewed by the human author, who takes responsibility for the document.

Acknowledgements

I thank collaborators in the Round Table research group for methodological discussions and conceptual feedback. The companion Moreau Arena project was shaped through iterative work with Claude, Gemini, Grok, GPT, Claude Code, and Codex as both research subjects and development tools. Any errors or misinterpretations in this paper are my own.

References

[1] V. Stasiuc. Moreau Arena: Not All LLMs Need Hints to Reason Strategically. Preprint, 2026.

[2] M. Mazeika et al. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[3] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *ACL*, 2022.

[4] A. Pan et al. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. In *ICML*, 2023.

[5] E. Awad et al. The Moral Machine experiment. *Nature*, 563:59–64, 2018.

[6] A. Bakhtin et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[7] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[8] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[9] KRAFTON AI. Orak: A Multi-Game Benchmark for LLM Agents. In *ICLR*, 2026.