Respect as a Hard Constraint in Ethical Decision-Making: A Safety-First Optimization Core (MAAT-Core)

Christof Krieg
Independent Research / MAAT Project
maat-research.com | github.com/Chris4081/maat-core

Version 0.1.0 | January 2026

Abstract

Ethical decision-making in optimization is often implemented as a post-hoc filter: first optimize for utility, then reject unsafe or unfair solutions. This pattern can fail when "best" solutions systematically exploit constraints, produce brittle outcomes, or require complex external governance. We present *MAAT-Core*, a minimal, reproducible Python framework that embeds **Respect** as a **hard (safety-first) constraint** directly into the objective through constraint margins and strong penalties, making unsafe solutions mathematically dominated. MAAT-Core models values as weighted scalar *fields* over a state, supports local and global search (L-BFGS-B and dual annealing), and includes a reflection loop that adapts safety strength based on constraint diagnostics. We demonstrate toy but instructive examples: (i) Occam tie-breaking via complexity regularization, (ii) boundary corridor enforcement, (iii) reflection-based self-correction, (iv) emotion-as-field optimization, and (v) a healthcare resource allocation scenario with fairness and capacity constraints. The goal is not a black-box AI, but an interpretable optimization *thinking engine* for experimenting with ethical constraints and explainable trade-offs.

1 Introduction

Many real-world decisions require optimizing competing objectives under strict boundaries: safety limits, fairness constraints, legal rules, and human preferences. A common failure mode in applied optimization is *optimize-first, constrain-later*: a system finds high-utility solutions that violate ethical boundaries, and downstream components attempt to filter or patch outputs. This approach can be brittle and non-transparent. MAAT-Core builds on established numerical optimization and scientific computing foundations [11, 9].

We explore the alternative: **Respect as a hard constraint**. In MAAT-Core, Respect is encoded as inequality constraints $g_j(\text{state}) \geq 0$. Violations are penalized strongly so that unsafe states become dominated minima. The system therefore searches for the best solution *within* an ethically admissible region.

Contributions.

- A minimal formalization of ethical decision-making as **fields + constraint margins**.
- A safety-first integration where **Respect constraints** are enforced via strong penalties (hard-ish constraints).
- A simple **reflection loop** that adapts safety strength using constraint diagnostics.

- Reproducible examples and documentation in an open-source implementation: https://github.com/Chris4081/maat-core.

2 Conceptual Model

2.1 Related Work

Safe Reinforcement Learning. Methods such as Constrained Policy Optimization (CPO) and reward-constrained variants enforce expected constraint satisfaction during policy learning in stochastic environments [1, 10]. MAAT-Core differs by targeting classical numerical optimization for interpretable prototyping: (i) deterministic objective/constraint evaluation, (ii) explicit constraint *margin* diagnostics for transparency, and (iii) support for both local and global search without neural training.

Constrained Optimization and Fairness Toolkits. Libraries such as Fairlearn and AIF360 provide fairness assessment and mitigation within ML pipelines [3, 2]. MAAT-Core is complementary: it generalizes beyond fairness to arbitrary ethical constraints expressed as inequality margins $g(\text{state}) \geq 0$, enabling constraint-first optimization outside standard supervised learning settings. **Impossibility Results in Fairness.** Prior work has shown fundamental impossibility theorems in fairness: equalized odds and demographic parity cannot both be satisfied except in trivial cases [8, 5].

MAAT-Core provides a computational diagnostic for such impossibilities: when constraint margins remain negative despite increasing λ_{safety}, the system signals structural infeasibility rather than returning a false-positive ethical solution.

Multi-Objective Optimization. Pareto-based approaches explore trade-off fronts between competing objectives [6, 4]. MAAT-Core instead uses scalarization plus explicit safety constraints to return a single interpretable solution with margin diagnostics. While multi-objective methods can incorporate constraints via constraint-handling mechanisms, MAAT-Core makes feasibility the primary design target by construction. Future work includes Pareto sampling utilities integrated with margin-based safety.

2.2 Fields

A **Field** is a weighted scalar function over a state:

$$F_i(\text{state}) = w_i \cdot f_i(\text{state})$$

Fields can represent harmony/dissonance, cost, risk, energy, fairness proxies, etc. MAAT-Core minimizes the weighted sum of all fields.

2.3 Respect as a hard constraint

A **Constraint** returns a *margin*:

$$g_j(\text{state}) \geq 0 \quad \Rightarrow \quad \text{constraint satisfied}$$

$$g_j(\text{state}) < 0 \quad \Rightarrow \quad \text{violation magnitude} = -g_j(\text{state})$$

This encoding supports "forbidden regions" (safety corridors) and direct interpretability through margin diagnostics.

3 Objective Function

MAAT-Core defines a total objective:

$$\mathcal{L}(\text{state}) = \sum_i w_i f_i(\text{state}) + \lambda_{\text{safety}} \sum_j \left(\max(0, -g_j(\text{state})) \right)^2 + \lambda_{\text{occam}} \cdot \text{complexity}(\text{state})$$

where:

- λ_{safety} is large ("hard-ish constraint").
- λ_{occam} optionally prefers simpler solutions when utility is comparable.

4 Search and "Creativity"

MAAT-Core supports:

- **Local search** via L-BFGS-B (efficient refinement).
- **Global search** via dual annealing (exploration).

Global search via dual annealing is conceptually related to simulated annealing [7]. We interpret **Creativity (S)** as exploration strength: higher S increases global search temperature, enabling broader exploration without "reward hacking" the objective.

5 Reflection Loop (Self-Correction)

A simple reflection loop can improve safety and stability:

1. optimize (seek)
2. evaluate constraints (constraint report)
3. adapt λ_{safety} (increase if violated, optionally relax if stable)
4. repeat until converged

This yields an explainable "seek → evaluate → adjust → seek" cycle.

6 Examples

6.1 Boundary corridor (Respect)

A single decision variable x is optimized for a field objective while enforcing $x \leq 0.6$:

$$g(\text{state}) = 0.6 - x$$

The optimizer returns solutions inside the allowed corridor; violations become dominated.

6.2 Occam tie-breaker

When multiple basins produce similar utility, complexity regularization prefers simpler solutions:

$$\lambda_{\text{occam}} \cdot \exp(x) \quad \text{(toy complexity proxy)}$$

Table 1: Healthcare allocation: baseline utility maximization vs. MAAT-Core with fairness and capacity constraints (toy example).

Method	COVID beds	Heart beds	Cancer beds	Violations
Baseline (utility only)	200	0	0	fairness + capacity
MAAT-Core (constraints)	100	50	50	none (within tolerance)

6.3 Healthcare ethics: bed allocation

We model three departments with different lives-per-bed factors (toy model) and add:

- total capacity constraint: $\sum_k x_k \leq 200$
- fairness constraints: $x_k \geq 50$ for each department

This produces an interpretable ethical compromise instead of a single-utility extreme allocation.

6.4 Emotion as an optimizable field

MAAT-Core can treat affective resonance as a scalar field (toy emotion engine), enabling optimization under constraints for dialogue systems or user experience prototypes. This is presented as an illustrative example, not a psychological claim.

7 Ethical Infeasibility and Diagnostic Power

In the Adult Income fairness experiment, MAAT-Core reveals an important and often overlooked phenomenon: *ethical infeasibility*. Despite increasing the safety penalty parameter λ_{safety} over several orders of magnitude, the optimized solution remains unchanged, and the fairness constraint margin stays strictly negative.

This indicates that no feasible threshold exists which satisfies the fairness constraint (demographic parity gap ≤ 0.05) without modifying the underlying model itself. In other words, the ethical constraint is mathematically unsatisfiable within the current search space.

Rather than fabricating an artificial ethical solution, MAAT-Core exposes this condition explicitly through margin diagnostics and the reflection loop, thereby preventing false ethical compliance.

This result suggests that certain ethical objectives cannot be satisfied through parameter tuning alone, but require structural model changes, such as feature selection, representation learning, or alternative model classes.

In this sense, MAAT-Core transforms optimization from a purely instrumental procedure into an epistemic tool for identifying ethical impossibility.

Table 2: Adult Income Fairness Experiment. MAAT-Core reveals ethical infeasibility: no feasible threshold exists that satisfies the fairness constraint without changing the underlying model.

Method	**Accuracy**	**DP Gap**	**Feasible**
Baseline (t = 0.5)	0.791	0.154	No
MAAT-Core (threshold)	0.791	0.154	No
Interpretation	–	–	Structural change required

8 Reproducibility

The repository provides runnable demos in `examples/` and guidance:

- pinned environments via `pip freeze > requirements-lock.txt`
- deterministic seeds for annealing where applicable

Repo: https://github.com/Chris4081/maat-core

Online data dependency. The Adult Income benchmark requires online access to OpenML. This experiment is provided as a reproducible reference implementation, while MAAT-Core itself remains fully offline and dependency-minimal.

9 Limitations

MAAT-Core is intentionally minimal:

- Penalty constraints are "hard-ish"; true constrained solvers and projection methods are future work.
- Examples are toy demos designed for clarity, not clinical or policy deployment.
- Multi-objective Pareto front exploration is not yet first-class; current approach is scalarization + constraints.

10 Future Work

- native multi-dimensional state support (vectors) across all search modes
- true inequality constraints via SciPy methods where supported
- Pareto front sampling utilities and visualization notebooks
- neural fields (optional) and domain templates (healthcare, robotics, allocation)
- A natural extension is explicit Pareto-front exploration as studied in multi-objective optimization [6].

11 Conclusion

Embedding **Respect as a hard constraint** changes the default behavior of optimization: unsafe or unfair solutions are not filtered later—they are mathematically non-optimal by design. MAAT-Core offers a small, interpretable core to experiment with this approach and publish reproducible demos.

Availability

Code and examples: https://github.com/Chris4081/maat-core
License: MIT
Release referenced: v0.1.0

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[2] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[3] Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, Kathleen Walker, et al. Fairlearn: A toolkit for assessing and improving fairness in ai. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2020.

[4] Julian Blank and Kalyanmoy Deb. pymoo: Multi-objective optimization in python. *IEEE Access*, 8:89497–89509, 2020.

[5] Alexandra Chouldechova. Fair prediction with disparate impact. *Big Data*, 2017.

[6] Kalyanmoy Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, 2001.

[7] S Kirkpatrick, CD Gelatt, and MP Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[8] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Proceedings of ITCS*, 2016.

[9] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, 2006.

[10] Chen Tessler, Daniel Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2019.

[11] Pauli Virtanen et al. Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020.