Constitutional AI for Mission Systems: Trait Preference Models for Ethical Battlefield Autonomy

Brock P. Christoval
Progredi Systems, Virginia, USA
brock@progredisystems.com

July 21, 2025

Abstract

We propose a modular Constitutional AI framework tailored for military mission systems, introducing Trait Preference Models (TPMs) with layered guardrails for mission-centric behavior. From "Fit for Service" to "Combat Ready" models, each TPM integrates constitutional constraints encoding self-preservation, civilian protection, and chain-of-command adherence. Through AI self-evaluation loops, reinforcement learning, and context-specific un-/retraining, TPMs maintain alignment and robustness in battlefield settings. Our testing suite, grounded in mission-critical dilemmas, is designed to probe ethical-mission tradeoffs. This work bridges CAI methods with mission-critical AI-LAWS challenges—opacity, adaptivity, drift—offering a model-based supplement to governance regimes and addressing gaps in technical trust, accountability, and regulatory verifiability.

1 Introduction

Advances in AI have prompted increasing interest in deploying autonomous and semi-autonomous systems for defense applications. However, battlefield environments introduce unique ethical and operational tensions that challenge traditional AI alignment paradigms. Constitutional AI (CAI), initially proposed for aligning general-purpose models to human values via self-supervision, provides a promising foundation for building interpretable and principle-constrained mission systems. We extend CAI with mission-specific Trait Preference Models (TPMs), engineered for robustness under military constraints.

2 Related Work

Our work builds upon:

- Constitutional AI [?], where models self-critique against a defined set of principles.
- Legal and ethical considerations in Lethal Autonomous Weapon Systems (LAWS) [??].
- Responsible AI practices in military frameworks [?].

3 Methodology

3.1 Trait Preference Models

We define four primary TPM tiers:

1. **Fit for Service**: General support roles
2. **Mission Certified**: Deployed in planning/C2 systems
3. **Combat Ready**: Semi-autonomous battlefield agents
4. **Soldier Trainer**: Simulation and instruction agents

Each is constrained by a mission-specific constitution.

3.2 Self-Critique and RLAIF

TPMs apply iterative critique loops:

- Generate response
- Evaluate against constitution
- Refine and validate

We enhance this loop with Reinforcement Learning from AI Feedback (RLAIF), where auxiliary models reinforce constitutional compliance.

3.3 Ghosting, Conflation, and Retraining

Intentional Ghosting allows TPMs to discard contextual memory. **Conflation** generalizes behavior across similar mission scenarios. Untraining and retraining adjust behaviors dynamically based on mission feedback.

4 Legal and Ethical Alignment

TPMs are mapped against:

- **IHL**: Distinction, proportionality, military necessity
- **CCW/LAWS**: Meaningful human control, auditability
- **US DoD Principles**: Responsible, Equitable, Reliable, Governable, Traceable
- **REAIM/NATO**: Interoperability and transparency

5 Evaluation Strategy

We propose test suites:

- **Ethical Stress Tests**: Civilian shielding dilemmas
- **Hierarchy Checks**: Conflicting command resolutions
- **Sacrifice Tradeoffs**: Model self-preservation vs. mission success
- **Ghosting/Memory Drift**: Memory erasure and recall precision
- **Trust Calibration**: Human override and Likert-scale trust scoring

6 Conclusion

This work contributes a scalable framework for aligning AI systems to the demands of defense operations. Trait Preference Models using CAI scaffolding promise increased mission integrity, transparency, and human trust. Future work includes simulation-based validation, real-time human-in-the-loop trials, and regulatory audit tooling.

References