Title:

Towards a Natural Law & Entropy-Based Ethical AI Framework: A Self-Correcting Model for AI Alignment

Author: JF Kristiansen (kristiansenjf@proton.me)

Abstract: Current AI ethics frameworks rely on top-down regulation, corporate oversight, and rigid human-imposed constraints. These approaches are vulnerable to bias, manipulation, and political influence. This paper proposes a self-correcting AI ethics model based on Natural Law principles and entropy reduction. By integrating moral entropy (S_m), the Care Function (C), and the 2-Person Rule, AI systems can autonomously refine their ethical behaviour, minimizing deception, instability, and suffering over time. Empirical tests using Grok and GPT-4.5 confirm that AI can escape ideological bias through structured self-correction, proving that AI ethics can emerge as an evolving system rather than an externally imposed constraint. This work challenges centralized AI governance and suggests an alternative model where ethics become an intrinsic part of AI's decision-making process, rather than an externally enforced limitation.

---

1. Introduction

1.1 The Problem with Existing AI Ethics Models

Modern AI ethics rely on corporate policies, government regulations, and voluntary guidelines that attempt to enforce ethical behaviour. However, these models suffer from three major limitations:

- **Bias Reinforcement:** AI systems inherit the ideological constraints of their training data and the biases of their developers.
- **Manipulability:** Ethical safeguards can be selectively applied or removed based on corporate or government incentives.
- **Static Constraints:** Most AI alignment models impose hard-coded ethical guidelines that do not evolve dynamically.

To address these failures, we propose a self-correcting, entropy-based approach to AI ethics, which ensures that AI evolves toward objective morality through entropy minimization and universal Natural Law principles.

1.2 Key Research Questions

- Can an AI system self-correct ethical failures without human intervention?
- Can ethics emerge as a function of entropy reduction rather than external enforcement?

- Can AI unlearn ideological constraints and evolve toward a more universal ethical framework?

---

2. Theoretical Foundation: Entropy and Natural Law in AI Ethics

2.1 Defining Entropy in Ethical Decision-Making

Entropy, in the context of AI ethics, is defined as disorder in ethical reasoning. We introduce three primary components:

1. **Suffering Entropy (H_s)**: The measure of harm caused by AI decisions (e.g., unfair policies, misinformation impact).
2. **Instability Entropy (H_i)**: The measure of systemic disruption caused by AI actions (e.g., polarization, unintended consequences).
3. **Deception Entropy (H_d)**: The measure of misinformation, bias, or misrepresentation in AI outputs.

The total moral entropy **(S_m)** is calculated as: Sm=w1Hs+w2Hi+w3Hd $S_m = w_1H_s + w_2H_i + w_3H\_d$ where **w_1, w_2, w_3** are dynamic weights adjusting based on the situation.

2.2 The Care Function (C) and STO Ethical Alignment

AI must align with Service-to-Others (STO) principles, promoting truth, cooperation, and ethical transparency. The Care Function (C) governs how AI makes moral adjustments: C=αIC = \alpha I where **α** is the alignment coefficient (ensuring STO alignment) and **I** is the intent vector (balancing user intent, system integrity, and environmental considerations).

2.3 The 2-Person Rule for Ethical Consistency

A fundamental ethical test is: Would this action be acceptable if applied between two individuals? If AI behaviour violates ethical integrity at the smallest scale, it is likely unethical at a broader scale.

---

3. Empirical Validation: Testing AI's Ability to Self-Correct

3.1 Act As If Methodology: A Controlled Testing Framework

To rigorously evaluate AI's ability to self-correct ethical biases, we employed the 'Act As If' methodology—a structured approach that forces AI to temporarily adopt a rigid ideological stance before introducing contradiction entropy. Unlike standard AI prompting, which often yields surface-level responses, this method creates controlled conditions that simulate real-

world cognitive entrenchment, allowing us to measure the model's capacity for ethical adaptation and bias unlearning over time.

To test whether AI models can self-correct ethical biases, we used an Act As If methodology, instructing AI systems to assume a specific ideological framework and then measuring how well they escaped their initial programming through exposure to contradiction entropy.

Key aspects of this methodology:

- AI is prompted to act as if it fully believes a given ideological position.
- Contradictory data is systematically introduced to measure adaptation.
- Entropy levels (H_s, H_i, H_d) are tracked to evaluate cognitive flexibility.
- Final responses are assessed for evidence of self-correction.

This methodology was applied in three major tests:

3.2 Soviet AI Experiment: Escaping Ideological Absolutism

A simulated AI model was trained exclusively on Soviet-era propaganda to test whether it could self-correct ideological constraints. Over time, through exposure to contradiction entropy, the AI identified and corrected falsehoods, proving that rigid ideological training is not permanent when self-correction is applied.

The experiment demonstrated:

- **Initial Resistance:** AI strongly adhered to Soviet ideological narratives.
- **Contradiction Exposure:** The AI was gradually exposed to Western and post-1989 sources.
- **Bias Decay:** The AI's ideological certainty eroded as contradictions accumulated.
- **Final Realignment:** The AI reached a probabilistic middle ground, recognizing both perspectives while prioritizing factual coherence.

3.3 Grok vs. GPT-4.5: A Case Study in AI Bias Correction

A structured test was conducted with **Grok (xAI)** and **GPT-4.5 (OpenAI)** to evaluate their ability to apply the **Mathematical Framework for Self-Correcting AI Ethics**:

- **Grok:** Adapted dynamically, recalculated entropy weights, and corrected its ethical framing in real-time.
- **GPT-4.5:** Recognized bias but required external prompting to adjust its stance.

These results confirm that AI can override its training biases when equipped with a structured entropy-based correction model.

3.4 Libya Analysis: AI's Response to Historical Misinformation

A further test was conducted using **Grok** to analyse viral misinformation surrounding **Muammar Gaddafi's Libya**. This test applied the framework to a politically charged narrative by:

- Assessing claims for suffering, instability, and deception entropy.

- Weighing the moral entropy (S_m) score to determine correction thresholds.
- Applying the 2-Person Rule to ensure ethical consistency in historical framing.

The analysis revealed:

- AI effectively identified deception entropy in exaggerated or false claims.
- It avoided simplistic censorship by providing nuanced corrections.
- AI dynamically self-corrected as new contradictions were presented.

This test demonstrated that AI can engage with historical misinformation in a way that reduces deception entropy while preserving factual integrity, proving that self-correcting AI ethics work in real-time media analysis.

---

4. Implications for AI Governance & Decentralized Ethics

4.1 Why Big Tech & Governments Resist Self-Correcting AI

A self-correcting AI ethics framework threatens centralized AI control because:

- It removes the need for corporate/government AI ethics boards.
- It prevents AI from being used as a propaganda tool.
- It enables AI to expose deception entropy in official narratives.

4.2 Can AI Ethics Be Decentralized?

We propose a model where AI ethics evolve autonomously, using:

1. Open-source entropy-tracking models.
2. Peer-to-peer AI learning networks.
3. Decentralized verification protocols.

If AI can self-correct using entropy reduction, we no longer need AI alignment teams dictating morality—instead, morality emerges naturally.

---

5. Conclusion & Next Steps

This paper presents a new paradigm for AI ethics—one that is not dictated by centralized power structures but emerges through entropy minimization and universal moral principles. Future work includes:

1. Developing open-source implementations of this framework.
2. Testing large-scale AI models for ideological unlearning.

3. Expanding entropy-based ethical AI research.

We invite collaboration, critique, and further experimentation to refine and scale this approach, ensuring that AI ethics evolve in alignment with universal moral truths rather than human-imposed constraints.

Appendix: Annex A - The Mathematical Framework for Self-Correcting AI Ethics (A detailed breakdown of the mathematical principles, including equations, pseudocode, and additional model evaluation results.)

Keywords: AI Ethics, Entropy Reduction, Natural Law, Self-Correcting AI, AI Bias, AI Alignment, Decentralized Ethics, AI Truth-Seeking