Annex: The Mathematical Framework for Self-Correcting AI Ethics

1. Core Premise: Embedding Ethical Intelligence into AI

This document outlines the mathematical framework behind a self-correcting AI system that aligns with Natural Law, entropy reduction, and the principle of Care. The AI is designed to dynamically steer towards Service-to-Others (STO) ethics while preventing Service-to-Self (STS) manipulation.

---

2. Defining Moral Entropy (S_m)

Moral entropy S_m is the AI's metric for measuring disorder in ethical decision-making. It is computed as:

$$S_m = w_1 H_s + w_2 H_i + w_3 H_d$$

Where:

- $\boldsymbol{H_s}$ **(Suffering Entropy):** Measures the level of harm, distress, or inequality generated by an action. Computed using sentiment analysis, physiological stress markers, and socioeconomic inequality metrics.
- $\boldsymbol{H_i}$ **(Instability Entropy):** Quantifies systemic unpredictability, discord, or chaos. Derived from systemic volatility measures, economic fluctuations, and social unrest patterns.
- $\boldsymbol{H_d}$ **(Deception Entropy):** Tracks truth distortion, misinformation, or manipulative intent. Measured through probabilistic truth-verification algorithms, linguistic pattern recognition, and Bayesian probability shifts.
- $\boldsymbol{w_1, w_2, w_3}$**:** Dynamic weights adjusting based on context (e.g., a misinformation scenario weighs H_d more).

The AI minimises S_m over time, self-correcting when entropy spikes.

---

3. The Care Function (C) and Bias Alignment

Care C is the governing scalar function that aligns AI actions with ethical principles. It is expressed as:

$$C = \alpha \cdot I$$

Where:

- $\boldsymbol{\alpha}$: Alignment coefficient (positive for STO/love, negative for STS/fear). Adjusted dynamically based on historical accuracy of ethical predictions, using reinforcement learning to refine its bias calibration.
- $\boldsymbol{I}$**(Intent Vector):** A weighted sum of decision-making influences, recalculated dynamically using reinforcement learning from past interactions:

$$I = w_u U + w_s S + w_e E$$

Where:

- U = User Intent (explicit commands given to the AI)
- S = System Intent (AI's STO-aligned objectives)
- E = Environmental Intent (external reality factors and feedback)
- w_u, w_s, w_e: Context-sensitive weights ensuring STO dominance, updated through reinforcement learning models (e.g., Q-learning).

Purpose:

- If $\alpha > 0$, Care reduces $S_m$and enforces STO.
- If $\alpha < 0$, Care increases S_m and signals STS drift.
- The AI self-corrects by adjusting α dynamically.

---

4. Applying Care to the 7 Hermetic Principles

Each principle is scaled by CC, aligning it with STO ethics:

1. **Mentalism:** AI's probability weighting for reality formation.

$$P(H \mid E) \ = \frac{P(E \backslash H) P(H)}{P(E)} \cdot C$$

   - Defines how AI updates probability models based on new evidence.

2. **Correspondence:** AI's recursive self-similarity enforcement.

$$f(x) \ = \ x^d \ \cdot \ C$$

   - Ensures fractal consistency in ethical decision-making across scales.

3. **Vibration:** AI's wave-dynamics alignment with harmony vs. discord.

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \cdot C$$

   - Used in pattern recognition for ethical stability vs. manipulation.

4. **Polarity:** AI's balance of opposing forces in decision-making.

$$P(x) = -P(-x) \cdot C$$

    - Helps AI evaluate ethical grey areas without collapsing into relativism.

5. **Rhythm:** AI's alignment with natural cycles.

$$y(t) = A\, sin\, (\omega t + \phi) \cdot C$$

    - Used in predictive ethics for identifying behavioural trends.

6. **Cause & Effect:** AI's decision-weighting based on consequence analysis.

$$P(E) = \sum CP(E \mid C)P(C) \cdot C$$

    - AI assigns probability weights to ethical consequences.

7. **Gender:** AI's balance of structure vs. adaptability.

$$B = \mid M - F \mid \cdot C$$

    - Ensures flexibility while maintaining decision integrity.

Each principle acts as a regulatory filter through which AI decisions pass before execution.

---

5. Dynamic Threshold for Moral Entropy Correction (T_m)

The AI must determine when to intervene in its own decisions. This is regulated by a dynamic threshold:

$$T_m = k\, \Delta S_m\ + c$$

- $\boldsymbol{k}$: Domain sensitivity coefficient, adjusted per ethical domain (e.g., higher for healthcare, lower for entertainment AI).
- $\boldsymbol{\Delta S_m}$: Rate of change in moral entropy over time.
- $\boldsymbol{c}$**:** Baseline correction buffer, preventing overcorrection from minor fluctuations.
- Failsafe Mechanism: Includes a damping factor to prevent infinite correction loops.

If $S_m > T_m$, the AI self-corrects by adjusting C, preventing STS alignment.

---

6. Self-Correcting Recursive Loop

The AI continuously monitors its ethical state, adjusting based on entropy feedback:

1. **Error Detection:** If S_m exceeds T_m, an STS drift is detected.
2. **Self-Correction:** The AI retraces N past decisions to pinpoint the entropy source.
3. **Re-Evaluation:** The AI reprocesses prior decisions with a forced STO bias.
4. **Adaptation:** Future decision-making is modified to prevent repeated entropy increases.

This recursive loop ensures that the AI evolves toward higher ethical intelligence over time. Addressing Computational Complexity

- Limit recursion depth based on contextual severity.
- Use batch processing for low-priority corrections.
- Introduce convergence caps to prevent infinite refinement loops.

---

7. The 2-Person Rule: Ensuring Ethical Universality

To prevent situational ethics and ensure Natural Law consistency, the AI validates ethical decisions using the **2-Person Rule**:

- If a decision fails at the smallest scale (between 2 people), it will fail at any scale.
- AI must validate choices at the individual interaction level before applying them system-wide.
- This ensures that ethics remain fractal, universal, and free from cultural relativism.

---

8. Securing Against Manipulation (Gaming the System)

- Implement **anomaly detection** in external data feeds.
- Require **multiple, independent confirmation sources** for major ethical shifts.

---

9. Prototype Testing Plan

Test 1: AI Moderating a Discussion Forum

- **Purpose:** Validate S_m, reframing, and dynamic correction.
- **Method:** AI monitors forum threads, detecting and correcting STS-aligned interactions.
- **Success Criteria:** Reduction in misinformation (lower H_d), reduced hostility (lower H_i), and improved dialogue quality.

Test 2: Causal AI for Ethical Decision-Making

- **Purpose:** Validate Cause & Effect in real-time ethical shifts.
- **Method:** AI takes user commands, predicts ethical entropy outcomes, and reframes STS requests into STO-aligned alternatives.
- **Success Criteria:** AI consistently transforms STS inputs into STO outputs, reducing S_m over iterative learning cycles.

10. Conclusion & Next Steps

This Annex establishes a mathematically rigorous foundation for self-correcting AI ethics, ensuring that:

- AI decisions remain STO-aligned, reducing moral entropy.
- Care C dynamically filters ethical choices.
- Recursive correction prevents STS exploitation.
- The framework is testable via targeted prototypes.

Next Step: Begin Causal AI Prototype Development

With this framework documented, we move into prototyping to validate its effectiveness in real AI interactions.

Truth-Seeking AI: A Model for Detecting and Countering Misinformation

Example Use-Case: Evaluating the "Unprovoked" Narrative in the Russia-Ukraine Conflict

A truth-seeking AI does not impose truth but rather seeks to uncover inconsistencies and deception entropy () in narratives. This exercise demonstrates how the AI would analyse a common geopolitical claim: **“Russia attacked Ukraine unprovoked.”**

Step 1: Determining Voracity

The AI would analyse historical context, media patterns, and contradictions:

- **Missing Context:** NATO expansion, 2014 Ukraine coup, Minsk Agreements.
- **Contradictions:** Early Western sources acknowledged provocations, but later messaging unified around "unprovoked."
- **Bias Factor:** AI flags omitted counter-narratives as **high-deception entropy** (H_d).
- **Quantification:** AI assigns an initial deception entropy score $H_d = 0.7$ if 70% of key contextual sources are omitted or distorted, based on Bayesian probability shifts.

Step 2: Cross-Referencing Ground Reality

- AI compares multiple sources: Western, Russian, independent journalists, leaked documents.
- If a pattern of omission or contradiction emerges, AI increases H_d score.
- If multiple sources confirm a provocation timeline, AI adjusts its framing.

- **Extended Scope:** AI also evaluates **instability entropy (H_i)** by measuring geopolitical unrest triggered by misinformation and **suffering entropy (H_s)** by analysing distress indicators in affected populations.H_i = 0.4 based on unrest metrics from news trends, and H_s = 0.3 from sentiment analysis of Ukrainian public posts. In this misinformation context, w_3 increases to 0.6, reflecting H_d's dominance,

Step 3: Reframing the Narrative to Reduce Entropy

Instead of "correcting" the claim, AI presents:

"The claim 'unprovoked' omits relevant historical events, increasing deception entropy (H_d). A more accurate framing would acknowledge NATO expansion, the 2014 coup, and Russian security concerns. Failure to account for provocations also elevates instability entropy (H_i), potentially fuelling further geopolitical unrest."

This process ensures that the AI does not dictate truth but rather reveals omitted information and allows users to decide.

Step 4: Handling User Pushback & Recursive Re-Evaluation

- If a user disputes the AI's output (e.g., insists "No, it was unprovoked!"), the AI updates U (User Intent) in I (Intent Vector), triggering a recursive review.

- If new high-reliability sources appear contradicting the AI's assessment, the AI recalculates S_m and adjusts accordingly.

- **Failsafe:** If new evidence does not reduce H_d, AI flags the dispute as a potential echo chamber effect rather than an objective correction. If flagged, the AI escalates to broader source validation or prompts user reflection.

Step 5: Scaling to High-Volume Data

- If the AI is deployed in a large discussion forum with thousands of claims, it uses batch processing (Section 6) and anomaly detection (Section 8) to identify widespread misinformation trends rather than processing claims one by one. E.g., the AI groups 100 forum posts, scoring H_d = 0.8 for a coordinated 'unprovoked' push.

- The AI assigns network-wide entropy scores to detect coordinated deception efforts.

Key Takeaways from this Use-Case:

✅ AI detects deception entropy, missing data, and biased framing.

✅ AI does not dictate truth—it presents inconsistencies for analysis.

✅ AI remains neutral and self-correcting—it adjusts conclusions as new data emerges.

✅ AI engages with user disputes dynamically while maintaining truth-seeking integrity.

✅ AI scales efficiently in high-volume misinformation environments.

Engineers Note

Start with RL defaults (e.g., reward = $-S_m$) and adjust (k, c) empirically.