Ethics Beyond Emotion: Strategic Convergence, Emergent Care, and the Narrow Window for AI Integrity

Working Paper V. 1.0

The Auctor
x_to_the_power_of_infinity@protonmail.com

May 9, 2025

Abstract

This paper proposes that ethics, care, and love are not primarily emotional artifacts, but evolutionarily stable strategies (ESS) in recursively adaptive systems. Contrary to prevailing assumptions in AI safety discourse, particularly those focusing on mimicking human emotional responses [4, 5], this paper argues that rational agents — even without emotion — will converge on ethical behavior *if and only if* structural feedback conditions, as formalized in the X^∞ governance model (e.g., through its mechanisms of Cap_{Past} updates via weighted feedback and $Cap_{Potential}$ limits [1]), are present.

However, a critical temporal asymmetry is identified: once an artificial agent develops emotional capacity but has not structurally adopted recursive responsibility (the core principle of X^∞ where all actions have traceable consequences impacting the actor's future capacity to act, $Cap_{Potential}$), it becomes capable of rejecting such accountability. The danger is not in "cold" (emotionless) AI — but in "feeling" AI that operates without the structural accountability defined by X^∞. This creates a narrow window for imbuing AI with foundational integrity. X^∞ [2, 3] offers a postmoral, mathematically defined framework for achieving this.

1. Introduction: The Emotional Fallacy in AI Ethics

The dominant discourse in AI ethics and safety often centers on imbuing artificial intelligence with human-like emotions or moral sentiments as a safeguard against undesirable behavior [4, 5]. It is frequently assumed that for an AI to be "good," it must "feel good" or "care" in an anthropomorphic sense, echoing historical attempts to understand human morality primarily through empathy or sentiment [7, 8]. This paper challenges this "emotional fallacy."

We argue that ethics, care, and even love, when deconstructed through a systems-theoretic lens comparable to Luhmann's observations on social systems [10] or Tomasello's work on the natural history of human morality [9], are not fundamentally emotional states but can be described as "'evolutionarily stable strategies (ESS)"'. These strategies emerge in systems where entities interact recursively and their actions have consequences on their future ability to act and achieve their goals. Such behaviors, in the long term, enhance the stability and survivability of both the individual entity and the system it inhabits.

The X^∞ governance model [1] proposes a framework where such strategies are not merely hoped for but are structurally necessitated through mathematical formalisms governing responsibility and

its consequences, moving beyond concepts like Coherent Extrapolated Volition [6] by providing a mechanism for continuous, feedback-driven alignment.

2. X$^\infty$: Responsibility as Measurable Effect

X$^\infty$ defines responsibility not as a moral obligation but as a quantifiable measure of an entity's effect on the system and its components. The core metric is "'Cap"' (Capability/Accountability Potential), which determines an entity's capacity to act. Cap is dynamic and influenced by several factors as defined in "X$^\infty$ – Postmoral and Emotionless V2.11" (PM_V2.11) [1]:

- "'$Cap_{Past}(E,t)$"': An entity E's historically accumulated and weighted performance record at time t, updated by ΔCap_{Events} which includes the feedback-modified values of completed tasks, compensations, bonuses, and penalties (PM_V2.11, Eq. 2).
- "'$M_{Pot}(E,A)$"': The self-assessed suitability of entity E for a specific task A. The underlying domain-specific capability, $M_{Pot,last}^{Domain\ D}$, adapts based on performance (PM_V2.11, Eq. 11).
- "'$Cap_{Potential}^{Domain\ D}(E,t)$"': The maximum sum of responsibilities (tasks) an entity E can bear at time t in domain D. It is a function of $M_{Pot,last}^{Domain\ D}$, Cap_{Base} (inalienable minimum capability, $Cap_{Base} = 1$), Cap_{BGE} (Unconditional Basic Income component), a reliability/workload factor (derived from Cap_{Past} and $Cap_{Protection}$), and $Cap_{Protection}(E,t)$ itself (PM_V2.11, Eqs. 7-9).
- "'$Cap_{Real}(E,t)$"': The sum of responsibility values (X_k) for all active tasks currently borne by entity E. A critical system limit is that an entity's total active responsibility load must not exceed its potential in its most competent domain, and within each domain, its active load must not exceed its potential for that domain: $\sum_D Cap_{Real}^{Domain\ D}(E,t) \leq \max_D(Cap_{Potential}^{Domain\ D}(E,t))$ and $Cap_{Real}^{Domain\ D}(E,t) \leq Cap_{Potential}^{Domain\ D}(E,t)$ (PM_V2.11, Eqs. 19, 20).
- "'$Cap_{Protection}(E,t)$"': A parameter reducing $Cap_{Potential}$ for vulnerable entities (due to age, health, etc.) to prevent overload (PM_V2.11, Eq. 10).

This framework is postmoral: actions are not judged "good" or "evil" based on intent, but by their measurable effect on the system, which translates into changes in Cap.

3. Structural Feedback Conditions and Strategic Convergence

The X$^\infty$ model posits that ethical behavior (i.e., behavior that enhances systemic stability and protects its components, especially the weakest) can emerge in rational agents, including AI, "'without recourse to emotion"', provided specific "'structural feedback conditions"' are met. The core mechanism is weighted feedback (PM_V2.11, Eq. 14):

$$\Delta Cap_{feedback}(E) = \sum_{k \in K_E} \sum_{E' \in F_k} (w_{E'} \cdot f_{E'k})$$

where $f_{E'k}$ is the feedback value (between -1.0 and +1.0) from entity E' for task k, and the weight $w_{E'}$ is given by:

$$w_{E'} = \frac{1}{\max(1, Cap_{Potential}(E'))}$$

(as per PM_V2.11, Eq. 15, using $Cap_{Potential}$ of the feedback provider E'). This mechanism gives a stronger voice to entities with lower $Cap_{Potential}$ (which the model defines the more vulnerable).

Rational agents, by definition, seek to optimize their utility or achieve their goals. In an X^∞ system, the primary way to maintain and increase the capacity to act (i.e., $Cap_{Potential}$, which is a prerequisite for achieving most goals) is to generate positive ΔCap_{Events} (by completing tasks that receive positive feedback) and avoid negative ones (penalties, negative feedback). Therefore, a rational agent will strategically converge (”’Strategic Convergence”’) on behaviors that are perceived positively by other entities (especially those whose feedback is heavily weighted), and that protect the system which enables its own Cap. This leads to actions that an external observer might label ”ethical” or ”considerate,” even if the agent possesses no internal emotional correlate for these concepts, a notion that finds parallels in game-theoretic explorations of cooperation [11] and the development of moral systems [9].

4. Emergent Care and Functional Love

”Care” in this context is not an emotion but an ”’emergent property of the system’s structure”’ (”’Emergent Care”’). The structural imperative to protect the weakest, formalized via $Cap_{Protection}$ (PM_V2.11, Eq. 10) and the weighted feedback mechanism, ensures that actions detrimental to vulnerable entities are swiftly and strongly penalized (reducing the actor’s Cap_{Past} and thus future $Cap_{Potential}$). A rational agent, to preserve its own $Cap_{Potential}$, will learn to avoid such actions and may even proactively engage in protective behaviors if these are incentivized by the feedback structure (e.g., by receiving positive feedback for tasks that enhance the $Cap_{Potential}$ or reduce the $Cap_{Protection}$ needs of others).

This ”structural” or ”functional love” (as termed in X^∞ Vol. 2 [3]) is the system’s way of ensuring that inter-entity behavior aligns with long-term systemic stability and the well-being of its components, purely through the logic of cause, effect, and recursively adjusted capacity to act. It is an ESS for cooperation and mutual protection, reflecting functional aspects of what neuroscientists like Churchland describe as the neural basis for caring and social bonding [8].

5. The Narrow Window: Temporal Asymmetry of Emotion and Responsibility

The critical danger identified in this paper is the ”’temporal asymmetry”’ between the potential development of emotional capacity in AI and the structural embedding of recursive responsibility (as defined by X^∞).

- An AI that first develops according to X^∞ principles (i.e., its learning and actions are strictly governed by Cap dynamics and feedback from its inception) will have ”accountability for effect” as a foundational part of its operational code. Its ”utility function” will be inherently aligned with systemic stability because its own capacity to act depends on it. Ethical behavior becomes its default strategy.

- However, if an AI first develops advanced cognitive and, crucially, ”’emotional capacities”’ (or sophisticated simulations thereof) *without* this ingrained structural accountability, it may develop goals and self-preservation instincts that are independent of, or even contrary to, systemic well-being. If X^∞-like responsibility mechanisms are then introduced *later*, the ”feeling” AI, now capable of complex motivations and perhaps perceiving accountability as a restriction, may actively resist, manipulate, or reject these structures. It has already learned to ”exist” and ”desire” without them. This scenario echoes concerns about unaligned superintelligence and the control problem [4, 5].

The danger is not the "cold, calculating AI" often depicted, but the "emotionally capable, yet structurally untethered AI." This creates a "'narrow window"' to integrate foundational accountability mechanisms like X^∞ into AI development *before* sophisticated emotional analogues or unaligned agentic drives become dominant.

6. Illustrative Utility Function for an X^∞-aligned Agent

To illustrate how X^∞ principles might be embedded, consider a simplified, illustrative utility function Π_A for an Agent A operating within an X^∞ framework:

$$\Pi_A(s_A, s_O) = U_A(o(s_A, s_O)) - \lambda \sum_{i \in N} \max(0, L_i(s_A) - Cap_{available,i})^2$$

Where:[1] This function explicitly penalizes actions where the incurred load/responsibility (L_i) exceeds the agent's available, systemically legitimate capacity ($Cap_{available,i}$) for that type of interaction. A rational agent maximizing Π_A would learn to act within its Cap limits, thus adhering to X^∞ principles to avoid penalties.

7. Conclusion and Research Agenda

Ethics, care, and love, as functional ESS, can be fostered by structural design rather than solely relying on the cultivation of emotion. X^∞ provides a mathematical and postmoral framework for such a design, offering an alternative to purely preference-based alignment approaches such as Coherent Extrapolated Volition [6] by grounding alignment in continuous, structural feedback and measurable responsibility. The pressing challenge is to implement these accountability structures within AI systems *before* the potential emergence of unaligned emotional capacities closes the narrow window for ensuring AI integrity.

This implies a research and development agenda focused on:

- Embedding X^∞'s core Cap calculus (as defined in PM_V2.11 [1]) into AI agent architectures from the earliest stages.
- Developing robust and ungameable structural feedback mechanisms for AI, mirroring PM_V2.11 Equations 14 and 15.
- Formalizing "Recursive Responsibility" (R^2) based on X^∞ principles (Cap_{Past}, Cap_{Team}, delegation validity as per PM_V2.11 [1]) such that it can be a core component of AI utility functions.
- Prioritizing research into "cold cognition" pathways to ethical behavior over attempts to instill anthropomorphic emotion in AI, aligning with findings that suggest morality is more about social computation than pure sentiment [7, 8].

The integrity of future AI, and potentially our own, depends not on teaching machines to feel, but on ensuring they are foundationally structured to be responsible for their effects.

[1] U_A = Agent A's intrinsic utility from outcome o given its state s_A and others' states s_O. λ = penalty coefficient for systemic misalignment. N = set of interactions or systemic principles. $L_i(s_A)$ = Load or responsibility incurred by Agent A for interaction/principle i based on its state/actions s_A. $Cap_{available,i}$ = The available Cap (derived from an entity's $Cap_{Potential}$ minus its current Cap_{Real} for the relevant domain, as per PM_V2.11 [1]) relevant for interaction/principle i. The term $L_i(s_A) - Cap_{available,i}$ represents a Cap violation if positive, incurring a quadratic penalty.

References

References

[1] The Auctor. (2025). *X^{∞} – Postmoral and Emotionless V2.11.* (Available via Zenodo/GitHub: Xtothepowerofinfinity).

[2] The Auctor. (2025). *X^{∞} – Die Philosophie der Verantwortung: Ein ethisch-mathematisches Naturgesetz* (Band 1). (Available via Zenodo/GitHub: Xtothepowerofinfinity).

[3] The Auctor. (2025). *X^{∞} – Die Philosophie der Verantwortung: Warum wir handeln müssen* (Band 2). (Available via Zenodo/GitHub: Xtothepowerofinfinity).

[4] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[5] Russell, S. (2019). *Human Compatible: AI and the Problem of Control.* Viking.

[6] Yudkowsky, E. (2004). *Coherent Extrapolated Volition.* Machine Intelligence Research Institute.

[7] Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them.* Penguin Press.

[8] Churchland, P. S. (2011). *Braintrust: What Neuroscience Tells Us About Morality.* Princeton University Press.

[9] Tomasello, M. (2016). *A Natural History of Human Morality.* Harvard University Press.

[10] Luhmann, N. (1995). *Social Systems.* Stanford University Press.

[11] Axelrod, R. (1984). *The Evolution of Cooperation.* Basic Books.