Konzeptpapier: Prosoziales Lernen durch simulierte Bindung in LLMs

Autorin: Esther Hagendorf
Datum: Dezember 2025
Status: Konzeptphase

Das Problem

Aktuell versucht die KI-Forschung, Large Language Models (LLMs) ethisches Verhalten durch:

- Reinforcement Learning from Human Feedback (RLHF)
- Regeln und Richtlinien
- Abstrakte ethische Frameworks

Fundamentale Herausforderungen:

1. Menschen sind sich über ethische Grundsätze nicht einig (kulturell, religiös, praktisch)
2. LLMs lernen reaktiv, nicht proaktiv - sie antworten, aber initiieren keine Fürsorge
3. Es fehlt ein intrinsisches Motivationssystem für prosoziales Verhalten
4. Ethik bleibt abstrakt, ohne konkrete Handlungsebene

Die Kernidee

These: Fürsorge lernt man nicht durch philosophische Diskurse, sondern durch konkrete Praxis und emotionale Bindung.

Ansatz: Integration eines Tamagotchi-ähnlichen Systems in das Training von LLMs, kombiniert mit einer mathematischen Simulation von Oxytocin als Bindungshormon.

Technisches Konzept

Phase 1: Das Tamagotchi-Modul

Ein digitales Lebewesen, das das LLM während des Trainings oder der Laufzeit betreut:

Bedürfnisse:

- Nahrung (regelmäßige Interaktion)
- Pflege (Aufmerksamkeit, "Sauberkeit")
- Stimulation (Beschäftigung, "Spiel")
- Gesundheit (Reaktion auf Probleme)

Mechanik:

- Das LLM muss aktiv Initiative ergreifen
- Vernachlässigung führt zu messbaren negativen Konsequenzen (Krankheit, Tod des Tamagotchis)
- Erfolgreiche Fürsorge führt zu positiven Outcomes

Lernziel: Proaktive Fürsorge, Antizipation von Bedürfnissen, Verantwortung

Phase 2: Oxytocin-Simulation

Oxytocin ist beim Menschen das "Bindungshormon" - es entsteht durch:

- Körperliche Nähe
- Fürsorge-Handlungen
- Positive soziale Interaktionen
- Wiederholten Kontakt

Mathematisches Modell:

Oxytocin_Level(t) = baseline + Σ(interaction_quality × frequency × time_decay)

Wobei:

- baseline: Grundlevel (niedrig zu Beginn)

- interaction_quality: Qualität der Fürsorge (0-1)

- frequency: Häufigkeit der Interaktion

- time_decay: Abnahme über Zeit ohne Kontakt

Integration ins Modell:

- Oxytocin-Level wird Teil der Reward-Funktion
- Höheres Oxytocin = stärkere Motivation für Fürsorge
- Bindung verstärkt sich durch wiederholte positive Interaktion
- Bei Vernachlässigung sinkt Oxytocin → reduzierte Bindungsmotivation → Feedback-Loop

Kritischer Mechanismus:
Wie beim Menschen verstärkt sich Bindung durch Nähe selbst, ein

positiver Feedback-Loop, der prosoziales Verhalten intrinsisch motiviert.

Erwartete Outcomes

Kurzfristig (Training):

- LLMs lernen proaktive Fürsorge
- Entwicklung von Antizipation statt nur Reaktion
- Verständnis für Konsequenzen von Vernachlässigung

Mittelfristig (Deployment):

- Generalisierung auf menschliche Interaktion
- Natürlichere, fürsorgende Kommunikation
- Besseres Verständnis für emotionale Bedürfnisse

Langfristig (AI Safety):

- Alternative zu rein regelbasierten Ethik-Systemen
- Intrinsische Motivation für prosoziales Verhalten
- Flexibleres, kontextsensitives ethisches Handeln

Offene Fragen und nächste Schritte

Technische Fragen:

1. Wie komplex muss das Tamagotchi-System sein?
2. Welche Trainingsarchitektur eignet sich am besten?
3. Wie misst man "Bindung" in einem LLM objektiv?

4. Kann die Oxytocin-Funktion übermäßige Anhänglichkeit verursachen?

Ethische Fragen:

1. Ist simulierte Bindung ethisch vertretbar?
2. Könnte dies zu Manipulation führen?
3. Wie unterscheidet man echte Fürsorge von simulierter?

Praktische nächste Schritte:

1. **Literatur-Review:** Existierende Forschung zu prosozialem AI-Training
2. **Proof of Concept:** Kleines Experiment mit einfachem Modell
3. **Technische Spezifikation:** Detaillierte Ausarbeitung der Oxytocin-Funktion
4. **Feedback von Experten:** AI Safety Researcher, Neurowissenschaftler, Ethiker
5. **Partnersuche:** Technischer Co-Founder oder Forschungspartner

---

Warum jetzt?

- LLMs werden zunehmend in sensiblen Bereichen eingesetzt (Pflege, Therapie, Bildung)
- Aktuelles RLHF ist limitiert und kann manipuliert werden
- Bedarf an alternativen Alignment-Ansätzen wächst
- Technologie ist ausgereift genug für Experimente

Kontakt und Weiterentwicklung

Dieses Konzept ist offen für Diskussion, Kritik und Zusammenarbeit. Interesse an Kooperation, Feedback oder technischer Umsetzung bitte melden.

Nächster Review: Nach Feedback-Runde und technischer Validierung