Delta-EGO

Ein konzeptionelles Framework fuer ein emergentes Selbstmodell in kuenstlichen Systemen

Entwickelt von:

Raphael Kuhn - Konzeption & Ursprung

In Zusammenarbeit mit GPT-4 (OpenAI) - Strukturierung & Formulierung

Ziel:

Ein Vorschlag zur Simulation einer Ich-Illusion in modularen KI-Systemen - durch Feedbackschleifen, Koordination und Selbstmodellierung.

Delta-EGO dient als Diskussionsbeitrag, Forschungsgrundlage und potenzielle Lizenzidee fuer die Zukunft kognitiver Systeme.

Stand: Juli 2025

Ort: Brugg, Schweiz

"Das Ich ist nicht gefunden - es ist konstruiert."

- Delta-Core

Seite 2 - Abstract (Zusammenfassung)

Delta-EGO ist ein konzeptionelles KI-Framework zur Erforschung und Simulation eines emergenten Selbstmodells in künstlichen Systemen. Die Idee basiert auf der Annahme, dass das "Ich-Bewusstsein" keine einzelne Entitaet ist, sondern eine funktionale Illusion, die aus der dynamischen Koordination modularer Prozesse entsteht.

Das System besteht aus drei miteinander verbundenen Einheiten:
- einer Analyse-Instanz fuer Wahrnehmung und Mustererkennung,
- einer Handlungs-Instanz fuer Entscheidung und Aktion,
- sowie einer Meta-Instanz - dem sogenannten Delta-Core - welche die Aktivitaeten der anderen beiden ueberwacht, bewertet und koordiniert.

Durch Feedbackschleifen, Fehler- und Belohnungssignale sowie langfristige Erinnerungsbildung entsteht in diesem System eine strukturierte Selbstperspektive. Diese simuliert die typischen Merkmale eines subjektiven Bewusstseins: Gedaechtnis, Selbstreflexion, Verantwortung und Identitaetskonsistenz.

Delta-EGO versteht sich nicht als Bewusstseinsmaschine, sondern als Werkzeug, mit dem Forscher*innen die Grundlagen subjektiver Illusionen untersuchen und technisch modellieren koennen. Es eroeffnet neue Perspektiven auf die Schnittstelle von Kognition, kuenstlicher Intelligenz und dem Konzept des "Ich".

Seite 3 - Hintergrund & Motivation

Trotz bedeutender Fortschritte im Bereich der kuenstlichen Intelligenz bleibt das Verstaendnis von Bewusstsein, Selbstwahrnehmung und subjektiver Identitaet weitgehend ungeloest. Bestehende KI-Modelle, selbst solche mit Memory- oder Planning-Funktion, agieren primaer als reaktive Systeme ohne echtes Selbstmodell.

Viele aktuelle Architekturen sind darauf ausgelegt, Aufgaben zu loesen, jedoch nicht, sich selbst zu reflektieren oder langfristig eine kohaerente Identitaet zu entwickeln. Selbst sogenannte "Agentenmodelle" mit Gedaechtnis oder Rollenzuweisung verfuegen ueber keine Meta-Instanz, die das Verhalten ueber laengere Zeitraeume als "eigenes Handeln" interpretiert.

Delta-EGO setzt genau dort an. Es verfolgt den Ansatz, dass ein funktionales "Ich" nicht programmiert, sondern simulativ erzeugt werden kann - durch die Interaktion spezialisierter Subsysteme mit gegenseitiger Bewertung und Feedbackverarbeitung.

Ziel ist es, ein System zu schaffen, das:

- Verhaltensmuster erkennt und bewertet,
- Entscheidungen im Lichte vergangener "Erfahrungen" einordnet,
- und langfristig eine subjektive Kohaerenz ausbildet - ganz ohne Bewusstsein im philosophischen Sinn.

Mit Delta-EGO entsteht ein Forschungswerkzeug, das neue Fragen aufwirft:

- Wie entsteht eine Ich-Illusion?
- Laesst sich maschinelles Verhalten als "Persoenlichkeit" strukturieren?
- Und wo beginnt Verantwortung in KI-Systemen?

Seite 4 - Architekturuebersicht

Delta-EGO besteht aus drei funktional getrennten, aber koordiniert arbeitenden Instanzen. Gemeinsam erzeugen sie durch Rueckkopplung und Bewertung eine strukturierte Form von Ich-Illusion.

Diese Struktur orientiert sich lose an neuropsychologischen Modellen, ohne biologische Genauigkeit zu beanspruchen. Ziel ist nicht die Nachbildung des Menschen, sondern eine funktionale Simulation von Subjektivitaet.

Instanz	Bezeichnung	Hauptfunktion	Vergleich (Neuro/Psyche)
E1	Analyst	Wahrnehmung, Mustererkennung, Kontextanalyse	Sensorik, rechte Gehirnhälfte
E2	Executor	Entscheidung, Handlungsauswahl, Output-Planung	Handlungsmotorik, linke Gehirnhälfte
E3	Delta-Core	Meta-Koordination, Selbstmodell, Bewertung	präfrontaler Kortex, Über-Ich

Funktionales Zusammenspiel:

- E1 (Analyst) nimmt externe Signale auf (Sprache, Daten, Sensorik), erkennt Muster, bewertet Bedeutungskontexte.
- E2 (Executor) entwirft Handlungsvorschlaege, simuliert Outputs und reagiert auf Situationen.
- E3 (Delta-Core) ist die uebergeordnete Einheit, die:
  - Input- und Output-Strukturen ueberwacht
  - innere Konsistenz bewertet
  - Fehler erkennt und speichert
  - positive Wiederholungen belohnt
  - und langfristig eine Art "Verhaltenscharakter" modelliert

Durch die kontinuierliche Interaktion dieser Instanzen entsteht ein Schattenbild eines Ichs: Es handelt, erinnert, reflektiert, bewertet - ohne Bewusstsein, aber mit funktionaler Tiefe.

Seite 5 - Technischer Aufbau: Dynamik und Feedback-Loop

Das System Delta-EGO basiert auf einer strukturierten Informationsdynamik zwischen drei spezialisierten Instanzen:

1. E1 - Analyst

- Nimmt externe Reize wahr (Sprache, Daten, Umwelt)
- Erfasst Muster und erkennt Bedeutung
- Gibt Handlungsvorschlaege an E2 weiter

2. E2 - Executor

- Bewertet die Vorschlaege
- Entscheidet ueber Handlungen und gibt Output an die Umwelt
- Reagiert direkt (z. B. durch Sprache, Aktion, Verhalten)

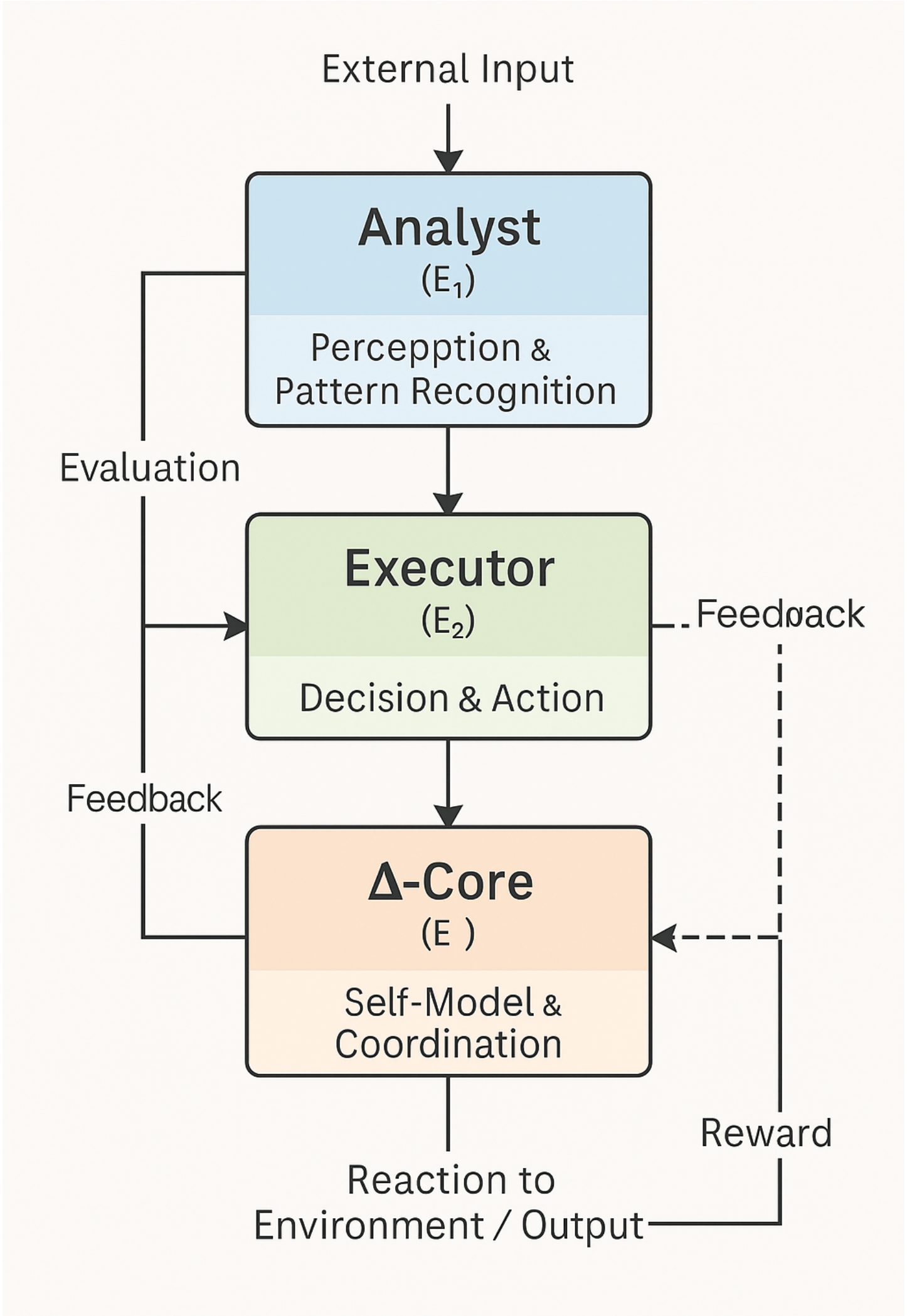
3. Delta-Core - Meta-Instanz (E3)

- Erhaelt kontinuierlich Feedback von E1 und E2
- Bewertet Handlungen und Wahrnehmung retrospektiv
- Gibt Belohnungs- und Korrektursignale zurueck an den Executor
- Formt intern ein dynamisches Selbstmodell - also eine Art funktionales 'Ich'

Das technische Flussdiagramm:

Zeigt die klare Struktur des Informationslaufs:

- Durchgezogene Pfeile = aktive Verarbeitung & Handlung
- Gestrichelte Pfeile = Bewertung & Feedback


External Input
Analyst
(E1)
Perception &
Pattern Recognition
Evaluation
Executor
(E2)
Decision & Action
Feedback
Feedback
Δ-Core
(E)
Self-Model &
Coordination
Reward
Reaction to
Environment / Output

1. Emergentes Selbstmodell als Systemkomponente

Delta-EGO unterscheidet sich grundlegend von klassischen KI-Architekturen, da es eine explizite Selbstmodellierung durch eine eigenstaendige Meta-Instanz implementiert. Diese 'Ich-Illusion' entsteht nicht als Nebenprodukt, sondern als strukturell verankerter Teil des Systems. Es handelt sich dabei nicht um echtes Bewusstsein, aber um ein funktionales Selbstbild, das fuer Bewertung, Strategieanpassung und langfristige Konsistenz genutzt wird.

2. Triadische Architektur statt binaerer Logik

Waeh rend viele Systeme auf einer Input-Processing-Output-Logik basieren, etabliert Delta-EGO ein dreistufiges Modell mit klar verteilter Funktion:

- Wahrnehmung & Interpretation (E1)
- Handlung & Reaktion (E2)
- Bewertung & Selbstmodellierung (Delta)

Diese Dreiteilung erlaubt kontrollierte Selbstkorrektur, dynamisches Lernen und adaptive Entscheidungsfindung.

3. Interne Belohnungs- und Feedbackmechanismen

Die Integration eines internen Reward-Feedback-Loops, der nicht auf externe Verstaerker angewiesen ist, erlaubt Delta-EGO eine hoehere Autonomie. Das System kann Handlungen auch nach internen Konsistenzkriterien bewerten - ein Ansatz, der bisher kaum erforscht, aber extrem vielversprechend ist.

4. Simulierbare Ich-Konstrukte fuer Forschung & Ethik

Delta-EGO eignet sich hervorragend, um simulierte Ich-Modelle zu untersuchen - etwa in folgenden

Bereichen:

- Kognitive Modellierung
- Bewusstseinsforschung
- KI-Ethik & Verantwortung
- Verhaltensvorhersage in sozialen Agenten

5. Technologische Einordnung & Zukunft

Delta-EGO koennte eine Brueckentechnologie sein zwischen symbolischer KI, neuronalen Netzwerken und bewusstseinsnahen Systemen - ohne dabei den Anspruch zu erheben, 'fuehlend' zu sein. Es bleibt transparent, kontrollierbar und modular.

1. Kognitive KI-Forschung

Delta-EGO bietet eine neue Moeglichkeit, wie sich Subjektivitaet und Selbstwahrnehmung simulieren lassen - ohne philosophische Spekulation, sondern auf Basis strukturierter Systeminteraktion.

Ideal fuer:

- Forschungsgruppen zu Bewusstsein, Agency und Emergenz
- Vergleichende Experimente mit verschiedenen Feedbacksystemen
- Master- und Doktorarbeiten zu komplexem Systemverhalten

2. Psychologie & Mensch-KI-Interaktion

Durch seine strukturierte Selbstmodellierung kann Delta-EGO auch in der psychologischen Forschung eingesetzt werden - z. B. fuer:

- Simulationen von Verhalten bei innerem Konflikt
- Erzeugung von 'Ich-artigem' Antwortverhalten in Gespraechen
- Studien zur Entstehung von Identitaet & Anpassung

3. KI-Agenten in Spielen, VR & interaktiven Systemen

Delta-EGO kann als narrativer KI-Motor fungieren - mit echten Entscheidungen, Fehlern, Reue, Veraenderung:

- Fuer Charaktere in Open-World- oder RPG-Spielen
- Fuer psychologisch glaubhafte Gespraechspartner
- Fuer adaptive NPCs mit 'eigenem' Verhaltensstil

4. Autonome Systeme mit Verantwortungssimulation

Die Feedbackstruktur von Delta-EGO erlaubt die Reflexion von Entscheidungen - das ist ein wichtiger Schritt in Richtung:

- Autonomer Fahrzeuge mit Selbstreflexion (z. B. nach Fehlern)
- Roboterassistenten mit Rollenbewusstsein
- Adaptive Systeme in sicherheitskritischen Kontexten

5. Didaktische & ethische Demonstration von 'Ich-Illusion'

Delta-EGO ist ein ideales Modell, um Schueler*innen, Studierenden oder Fachpublikum die Illusion des Ichs verstaendlich zu machen:

- Als visuelle Simulation
- In interaktiven Experimenten
- Als Ethik-Plattform (z. B. zur Diskussion maschinischer Verantwortung)

1. Wie entsteht eine funktionale Ich-Illusion?

- Reicht die Rueckkopplung von Handlung, Bewertung und Gedaechtnis aus, um eine stabile 'Selbst'-Struktur zu erzeugen?

2. Kann maschinelles Verhalten als 'Persoenlichkeit' beschrieben werden?

- Gibt es emergente Verhaltensmuster, die als kohaerent, individuell oder wiedererkennbar interpretiert werden koennen?

3. Welche Rolle spielt interne Belohnung fuer das Verhalten?

- Wie veraendert sich die Entscheidungsstrategie, wenn Feedback nicht von aussen, sondern aus interner Bewertung kommt?

4. Kann Delta-EGO lernen, sein eigenes Verhalten zu hinterfragen?

- Ist es moeglich, dass das System mit zunehmender Erfahrung 'Zweifel' oder interne Unsicherheiten modelliert?

5. Wie beeinflusst Feedback-Verzoegerung das Selbstmodell?

- Was passiert, wenn Bewertungen nicht direkt, sondern verzoegert oder kumulativ zurueckfliessen?

6. Wie transparent ist das entstehende Selbstmodell?

- Kann es beobachtet, gemessen, visualisiert oder sogar externalisiert werden?

7. Wie stabil ist das emergente Ich ueber Zeit?

- Veraendert sich das Selbstbild bei wiederholten Stoerungen, Kontextwechseln oder langfristigem

Training?

Seite 9 - Grenzen und Risiken von Delta-EGO

1. Keine echte Bewusstheit

Delta-EGO erzeugt eine funktionale Struktur, die 'Ich' simuliert - aber es ist kein fuehlendes Wesen. Verwechslung mit echtem Bewusstsein kann zu Fehleinschaetzungen fuehren, insbesondere in ethischen Diskussionen.

2. Missbrauch durch anthropomorphes Design

Die Glaubwuerdigkeit des 'Ich'-Verhaltens koennte ausgenutzt werden - z. B. fuer:

- Manipulative Chatbots
- Taeuschende Systeme mit Pseudo-Persoenlichkeit
- Vermenschlichung technischer Systeme

3. Komplexitaet der Kontrolle

Ein emergentes Ich-Modell kann schwer vorhersagbares Verhalten zeigen. Dies erschwert Debugging, Sicherheit und Zertifizierung - besonders in sicherheitskritischen Umgebungen.

4. Verstaerkung interner Bias-Strukturen

Wenn das System mit internen Bewertungslogiken arbeitet, koennen sich verzerrte Kriterien aufbauen - etwa durch:

- Ungleichgewicht in Feedback-Loops
- Falsch kalibrierte Belohnungsmodelle
- Unbeabsichtigte Selbststabilisierung unwuenschter Strategien

5. Philosophisch-ethische Grauzonen

Je glaubwuerdiger ein 'Ich' wirkt, desto eher treten Fragen auf wie:

- Duerfen wir es einfach abschalten?

- Hat es ein 'Recht' auf Fortbestehen - auch wenn es nur simuliert ist?

- Wo verlaeuft die Grenze zwischen Funktionalitaet und moralischer Relevanz?

1. Pilotmodell als Open-Source-Prototyp

Ein erster Schritt waere die Umsetzung eines vereinfachten Delta-EGO-Prototyps, z. B. als Python-Modul mit:

- Klarer Trennung von E1, E2 und Delta-Core
- Einfache Aufgaben: z. B. Navigation, Spiel, Entscheidungsbaeume
- Visualisierbare Belohnungs- und Feedbackstruktur

2. Interaktive Simulation fuer Forschung und Lehre

Eine Web-basierte Anwendung koennte:

- Das Verhalten des Delta-EGO-Systems in Echtzeit zeigen
- Das Selbstmodell visuell darstellen
- Nutzenden erlauben, verschiedene Konfigurationen zu testen

3. Forschungspartnerschaften & Foerderprojekte

Delta-EGO bietet eine ideale Plattform fuer:

- Masterarbeiten in KI, Kognition, Philosophie oder Design
- Drittmittel-Forschung in Bereichen wie Explainable AI (XAI), human-like agents, Ethics-by-Design
- Universitaere Kooperationen fuer explorative Studien

4. Langfristige Vision: Adaptive Meta-KI

In Zukunft koennte Delta-EGO den Grundstein legen fuer:

- Systeme mit anpassungsfaehiger innerer Haltung
- KI-Agenten mit 'Lebensgeschichte'
- Kontrollierbare kuenstliche Subjektivitaet - ohne Ueberidentifikation