Anthropic, „Claude Gov“ und die neue Verfassung – Eine ordnungspolitische Einordnung

Posted on Februar 1, 2026 by Redaktion-AnalyseTeam

Anthropic ist ein US-amerikanisches Unternehmen, das seit seiner Gründung 2021 an Sprachmodellen der *Claude*-Familie arbeitet und sich selbst als auf *sichere KI* ausgerichtet bezeichnet. Die Organisation ist als *Public Benefit Corporation* strukturiert, was bedeutet, dass neben kommerziellen auch soziale und sicherheitsbezogene Zielsetzungen im Firmenstatut stehen.

1. Claude Gov: KI für nationale Sicherheit

Bereits im Juni 2025 stellte Anthropic speziell angepasste Modelle namens **„Claude Gov“** vor, die ausdrücklich für nationale Sicherheitsbehörden entwickelt wurden. Diese Modelle sind laut Anthropic für Aufgaben wie Analyse klassifizierter Dokumente oder komplexe Dateninterpretation geeignet und werden in streng kontrollierten Umgebungen („classified environments“) eingesetzt.

Anthropic arbeitet in diesem Bereich mit Partnern wie **Palantir** zusammen, um KI-Fähigkeiten in Geheimdienst- und Verteidigungsnetzwerke zu integrieren. Dieser Einsatz beschränkt sich offenbar auf **Datenanalyse und Unterstützungsfunktionen**, nicht auf autonome Entscheidungssysteme.

2. Die neue Verfassung vom 22. Januar 2026

Am 22. Januar 2026 veröffentlichte Anthropic eine überarbeitete **„Verfassung“ für Claude**, die nicht nur Verhaltensregeln, sondern auch eine philosophische Rahmung von Werten und Prinzipien umfasst. Ziel ist, dass die KI nicht nur *weiß*, was sie tun soll, sondern *versteht*, warum bestimmte Verhaltensweisen gewünscht sind und andere nicht. Die Offenlegung dieses Dokuments unter einer CC0-Lizenz bedeutet zusätzlich, dass diese Leitprinzipien öffentlich einsehbar und nutzbar sind.

Diese Verfassung ist Teil von Anthropic’s Ansatz, KI-Ausrichtung **nicht nur technisch, sondern auch normativ transparent** zu gestalten. Sie soll dabei helfen, Werte und Abwägungen im KI-Training zu verankern und macht den ethischen Rahmen für Claude-Antworten nachvollziehbar.

3. Konflikt mit dem Pentagon

Ende Januar 2026 berichteten mehrere Medien, darunter Reuters und das *Wall Street Journal*, über wachsende Spannungen zwischen dem US-Verteidigungsministerium (Pentagon) und Anthropic. Der Kern des Konflikts betrifft **ethische Einschränkungen**, die Anthropic an seine Technologie bindet:

- Anthropic möchte verhindern, dass seine Modelle **autonome Waffensysteme targetieren** oder **zur Überwachung von US-Bürgern ohne menschliche Kontrolle** genutzt werden.
- Pentagon-Vertreter hingegen vertreten die Auffassung, dass kommerzielle KI dort eingesetzt werden sollte, wo es die **US-Gesetze erlauben**, auch wenn dies den firmeneigenen Nutzungsrichtlinien widerspricht.

Dieser Streit hat einen Vertragswert von mehreren hundert Millionen Dollar und könnte laut Berichten die weitere Kooperation mit dem Verteidigungsministerium gefährden.

4. Nutzung durch Strafverfolgungsbehörden

Die bisherigen öffentlich zugänglichen Quellen weisen nicht darauf hin, dass Anthropic explizit externen Behörden wie **FBI, Secret Service oder ICE** freie Nutzung ohne Einschränkungen gestattet hat. Der Streit mit dem Pentagon deutet im Gegenteil darauf hin, dass Anthropic derzeit **gerade solche Anwendungen einschränkt**, um Überwachung ohne menschliche Kontrolle auszuschließen.

Konkrete Details über nationale Polizeieinsätze oder gesetzliche Streitigkeiten im Weißen Haus liegen in den vertrauenswürdigen Berichten derzeit nicht vor.

5. Anthropics Position und Unternehmensstrategie

Anthropic hält an einem Ansatz fest, der gewisse **Einsatzbeschränkungen für KI-Technologien** vorsieht, insbesondere dort, wo KI autonom Entscheidungen treffen würde, die Menschen betreffen. Der CEO Dario Amodei hat öffentlich betont, dass KI die nationale Verteidigung unterstützen soll, „**außer in jenen Bereichen, die uns unseren autokratischen Gegnern ähnlicher machen würden**“.

Dieser Ansatz steht für das Unternehmen nicht im Widerspruch zu seiner Interessenlage – im Gegenteil: Er ist Teil seiner **Markenidentität als „sichere KI“**, die nicht nur kommerziell, sondern auch ethisch tragfähig sein will.

6. Bewertung im ordnungspolitischen Kontext

Aus ordnungspolitischer Sicht wirft dieser Konflikt zwei grundsätzliche Fragen auf:

1. **Wer legt die Einsatzbedingungen für leistungsfähige KI fest?**
   Ist es allein das Unternehmen, das sie entwickelt, oder muss es staatlichen oder internationalen Aufsichten unterliegen?
2. **Welche Verantwortung haben KI-Entwickler gegenüber dem Einsatz in militärischen oder geheimdienstlichen Kontexten?**
   Wenn ein Unternehmen Nutzung einschränkt, um menschenzentrierte Kontrolle zu gewährleisten, widerspricht das nicht zwangsläufig den staatlichen Interessen, kann aber zu Spannungen führen.

Anthropic befindet sich hier an einem Schnittpunkt:

- einerseits will es **breite Anwendung seiner Technologie**,
- andererseits setzt es **ethische Grenzen**, die selbststaatlichen Nutzungsansprüchen entgegenstehen.
  Das ist kein Zufall, sondern Ausdruck eines Unternehmenskonzepts, das **normative Orientierung im Technologievertrag** sucht statt unbeschränkter technokratischer Verfügbarkeit.

Fazit

Anthropics neue Verfassung ist nicht nur ein Dokument für Entwickler, sondern ein **öffentliches Bekenntnis zu bestimmten Grundwerten und Grenzziehungen** im KI-Kontext. Die damit verbundenen Spannungen mit staatlichen Akteuren wie dem Pentagon zeigen, dass KI-Governance nicht nur technisch, sondern auch politisch und ordnungspolitisch entschieden wird. In diesem Konflikt ist nicht allein KI das Thema, sondern **die Frage, wer über Zwecke, Grenzen und Verantwortlichkeiten entscheidet**.

Grafik: ChatGPT

Claudes neue Verfassung: https://www.anthropic.com/news/claude-new-constitution