GOVERNANCE & RESPONSIBLE-AI-MODELL · EDITION 2026.02

Autonomie, die Sie steuern können, ist die einzige Autonomie, die ein Deployment wert ist.

Dieses Dokument spezifiziert, wie ZeroMan.ai Autonomie begrenzt, auditiert und verdient - heute für die eigenen Agenten und in der Marketplace-Ära für Drittmodule. Es ist so geschrieben, dass man uns daran messen kann: Jede Verpflichtung hier ist im Produkt testbar.

POSITIONIERUNG

Volle Autonomie ohne Governance ist ein Risiko. Empfehlungen ohne Ausführung sind Theater. Die einzige glaubwürdige Enterprise-KI ist ein System, das innerhalb eines expliziten Entscheidungsrechte-Rahmens handelt - und dies nachweisen kann.

1 · Prinzipien

#	Prinzip	Verpflichtung
01	Begrenzte Autonomie	Kein Agent und kein Modul handelt außerhalb deklarierter Entscheidungsrechte, Schwellenwerte und Datenumfänge.
02	Menschliche Verantwortung für Strategie	Menschen besitzen Ziele, Richtlinien, Schwellenwerte und Ausnahmen. Das System besitzt die Koordination.
03	Erklärbarkeit von Anfang an	Jede Empfehlung enthält Annahmen, bindende Restriktionen und Zielkonflikte - mit der Entscheidung erzeugt, nicht nachträglich rekonstruiert.
04	Prüfbarkeit	Jeder Entscheidungslauf erzeugt einen unveränderlichen, wiederholbaren Datensatz. Was nicht protokolliert ist, ist nicht passiert.
05	Reversibilität & Override	Menschen können jede automatisierte Aktionsklasse jederzeit pausieren, übersteuern oder zurückrollen - und Overrides werden selbst protokolliert und zum Lernen genutzt.
06	Verhältnismäßigkeit	Die Autonomie einer Entscheidungsklasse richtet sich nach gemessener Zuverlässigkeit und begrenztem Auswirkungsradius - niemals nach Bequemlichkeit.

2 · Autonomiestufen

Stufe	Name	Verhalten	Typischer Einsatz
L1	Empfehlung	Analysiert, erklärt, schlägt vor. Keine Aktion.	Neue Schleifen; geringe Datenzuverlässigkeit.
L2	Freigabe zur Ausführung	Das System bereitet die Aktion vor; ein benannter Mensch genehmigt vor der Ausführung.	Standardhaltung für alle neuen Deployments.
L3	Begrenzte Autonomie	Führt automatisch innerhalb von Richtlinien, Schwellenwerten und Auswirkungsgrenzen aus.	Bewährte, risikoarme Aktionsklassen.
L4	Enterprise-Autonomie	Governance-gesteuerte Schleifen laufen Ende zu Ende; Menschen besitzen Strategie, Richtlinien und Ausnahmen.	Langfristiges Ziel, pro Domäne verdient.

WIE AUTONOMIE VERDIENT - UND VERLOREN - WIRD

Beförderung: Eine Entscheidungsklasse steigt nur dann um eine Stufe auf, wenn N aufeinanderfolgende Läufe ohne materielle Korrektur, innerhalb definierter Fehlergrenzen und über einen Mindestkalenderzeitraum genehmigt wurden - alle drei Parameter werden pro Klasse mit dem Kunden festgelegt. Herabstufung: Jeder Schwellenwertverstoß, jede materielle Korrektur oder Anomalie stuft die Klasse automatisch um eine Stufe zurück und öffnet eine Prüfung. Auswirkungsgrenzen (Wert-, Umfangs- und Ratenlimits) gelten auf jeder Stufe oberhalb von L1.

3 · Entscheidungsrechte und Genehmigungsmatrix

Entscheidungsrechte definieren, wer welche Aktionsart unter welchen Bedingungen genehmigen darf. Die folgende Matrix ist ein Referenzmuster; jedes Deployment instanziiert während der Design-Partnerschaft seine eigene Version.

Entscheidungstyp	Genehmigungslogik	Kontrollklasse
Anpassung der Nachschubmenge	Unterhalb der Wertschwelle automatisch erlaubt	Autonom (L3-fähig)
Änderung einer Bestellung	Erfordert Genehmigung oberhalb der Kostengrenze	Kostenschwelle
Änderung des Produktionsplans	Erfordert Genehmigung bei Kundenauswirkung	Kundenauswirkung
Bestandsumverteilung	Innerhalb der Richtlinie erlaubt; markiert bei regionsübergreifender Umverteilung	Richtliniengebunden
Frachtbeschleunigung	Erfordert Genehmigung oberhalb der Beschleunigungskostenschwelle	Kostenschwelle
Kundenzuteilung	Erfordert Genehmigung, wenn strategische Kunden betroffen sind	Kundenauswirkung
Lieferantenersatz	Erfordert Genehmigung bei Vertrags- oder Compliance-Risiko	Richtliniengebunden
Finanzieller Zielkonflikt	Executive-Genehmigung, wenn der Margen-/Service-Zielkonflikt die Richtlinie überschreitet	Executive

4 · Der Audit-Rahmen

Jeder Entscheidungslauf - Empfehlung oder Ausführung, menschlich genehmigt oder autonom - schreibt einen vollständigen, unveränderlichen Telemetriedatensatz:

```
TELEMETRIEDATENSATZ - PFLICHTFELDER
decision_id · loop_class · trigger_signal · severity
state_snapshot_ref · constraints[] · binding_constraints[]
agents_invoked[] · models_used[] · assumptions[]
scenarios[] · tradeoffs · recommendation · rationale
governance: rights_applied · thresholds_checked · approver · timestamp
actions_prepared[] · execution_channel · execution_status
outcome: realized_metrics · variance · decision_quality_score
overrides[] · corrections[] · review_flags[]
```

Datensätze werden gemäß Kundenrichtlinie aufbewahrt, auf Anfrage exportiert und so gestaltet, dass interne Audit-, Risiko- und Compliance-Prüfungen ohne Rekonstruktion möglich sind.

5 · Override, Eskalation und Stoppschalter

- Pause: Autorisierte Rollen können jede Entscheidungsklasse, jede Schleife oder die gesamte Plattform sofort pausieren; vorbereitete laufende Aktionen werden gehalten und nie halb ausgeführt.

- Override: Genehmigende können jede vorbereitete Aktion ändern oder ablehnen; Override, Begründung und Ergebnis gehen in die Lernschleife ein.

- Eskalation: Ungelöste Genehmigungen eskalieren zeitgesteuert durch benannte Ebenen (Owner -> Director -> Executive); Schweigen bedeutet niemals Zustimmung.

- Stoppschalter: Ein harter Kill pro Integrationskanal stellt sicher, dass bei Aktivierung keine Schreibvorgänge ein System of Record erreichen.

6 · Governance im Marketplace-Zeitalter

Wenn ZeroMan.ai sich für Drittmodule öffnet, gilt dieselbe Disziplin für das Ökosystem:

- Datenumfangs-Manifeste: Jedes Modul deklariert genau, was es liest und schreibt - app-artige Berechtigungen, vom Core durchgesetzt und für den Kunden sichtbar. Kein Umfang, keine Daten.

- Zertifizierung: Module bestehen vor dem Listing Konformitätstests (Verträge eingehalten, Telemetrie ausgegeben, Fehlermodi behandelt). Zertifizierung ist erneuerbar, nicht dauerhaft, und wird über Loop Zero, ZeroMans agentengeführtes Onboarding, administriert.

- Isolation zwischen Anbietern: Kein Modul sieht Daten, Prompts, Modelle oder Telemetrie eines anderen Anbieters. Isolation wird durch die Plattform erzwungen, nicht durch Richtliniendokumente.

- Gemessene Ratings, veröffentlichte Methodik: Anbieter-Scorecards werden aus Orchestrierungstelemetrie gegen Ground-Truth-Ergebnisse berechnet; die Methodik ist öffentlich, Scores sind nicht käuflich, und Anbieter haben ein dokumentiertes Einspruchsverfahren.

- Die Neutralitätsregel: ZeroMans eigene Module werden ausschließlich auf den öffentlichen APIs gebaut, die jedem Drittanbieter zur Verfügung stehen - keine privaten Hooks, keine privilegierten Daten, kein bevorzugtes Ranking. Der Schiedsrichter spielt nicht mit Heimvorteil.

7 · Verpflichtungen für verantwortungsvolle KI

- Kundendaten werden nicht zum Training gemeinsamer oder Foundation-Modelle verwendet; Lernen ist mandantenspezifisch, sofern nicht ausdrücklich anders vereinbart.

- Modell- und Annahmen-Nachvollziehbarkeit für jede Empfehlung; keine nicht zuordenbaren Entscheidungen.

- Entscheidungen mit menschlicher Wirkung (z. B. Personaleinsatzplanung, Lieferantenkündigung) bleiben standardmäßig L1/L2, unabhängig von gemessener Leistung.

- Bias- und Drift-Monitoring für Prognose- und Allokationsmodelle mit dokumentiertem Prüfzyklus.

- Materielle Automatisierungsvorfälle werden betroffenen Kunden mit Ursache und Abhilfe offengelegt.

Wir veröffentlichen dieses Dokument und laden Kunden, Partner und Forschende ein, uns daran zu messen.

8 · Umfang

ZeroMan.ai befindet sich im Frühstadium; die Plattform wird entworfen und gebaut. Dieses Dokument spezifiziert das Governance-Modell, gegen das die Plattform gebaut wird - es erhebt keine Ansprüche auf Zertifizierungen, Kunden oder Produktionsdeployments. Fragen, Kritik und Red-Team-Interesse sind willkommen: zeroman.ai/contact.