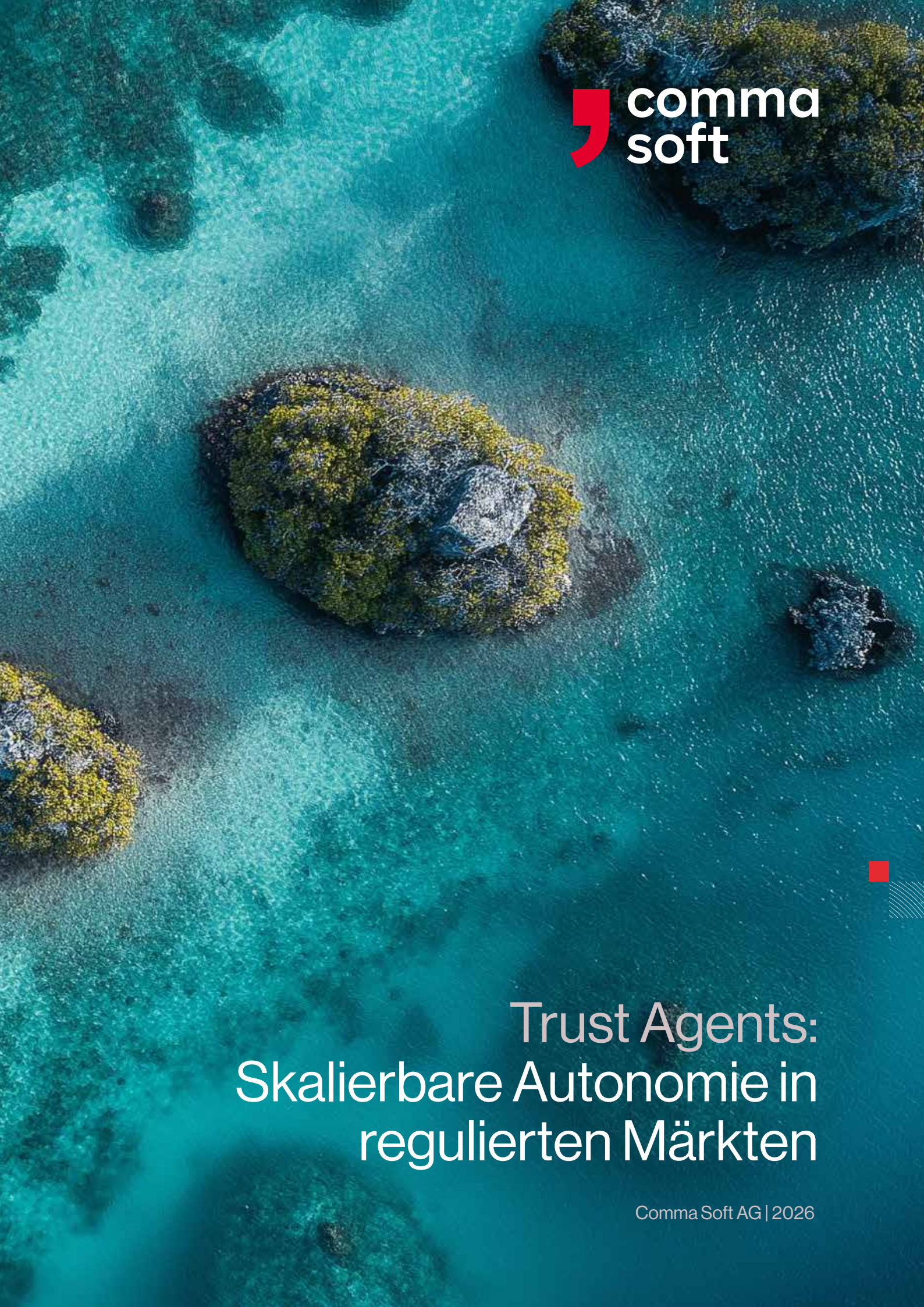


comma soft

Trust Agents: Skalierbare Autonomie in regulierten Märkten

Comma Soft AG | 2026

Inhaltsverzeichnis

1. Einführung & Executive Summary: Trust als Grundlage autonomer KI-Systeme

Künstliche Intelligenz entwickelt sich derzeit von einem unterstützenden Werkzeug zu einer eigenständig handelnden digitalen Instanz. Mit dem Aufkommen von KI-Agenten beginnt eine neue Phase der Unternehmensdigitalisierung: Systeme analysieren nicht mehr nur Informationen oder generieren Inhalte, sondern treffen Entscheidungen, koordinieren Prozesse und interagieren autonom mit Kunden, Mitarbeitenden und anderen Systemen.

Insbesondere in regulierten Branchen wie Versicherungen und Banken, im Gesundheitswesen, im Bereich Pharma, öffentlicher Verwaltung sowie Defence & Intelligence entsteht daraus ein erheblicher strategischer Hebel. KI-Agenten versprechen deutliche Effizienzgewinne, schnellere Prozesse, höhere Skalierbarkeit und eine neue Qualität operativer Automatisierung.

Gleichzeitig verschiebt sich jedoch die zentrale Managementfrage fundamental. Im Fokus steht nicht länger primär, ob Organisationen technisch in der Lage sind, KI-Agenten zu entwickeln. Die entscheidende Frage lautet vielmehr:

Wie lassen sich autonome KI-Systeme kontrollierbar, regelkonform und vertrauenswürdig betreiben?

Denn mit zunehmender Autonomie entstehen neue Risiken:

- Entscheidungen werden dynamischer und schwerer vorhersehbar,
- Verantwortlichkeiten diffuser,
- regulatorische Anforderungen komplexer,
- und klassische Governance-Modelle stoßen an ihre Grenzen.

Während Mitarbeitende über Richtlinien, Kontrollmechanismen, Eskalationswege und Governance-Strukturen gesteuert werden, fehlen vergleichbare operative Steuerungsmodelle für KI-Agenten bislang weitgehend. Genau hier entsteht aktuell eine der größten strategischen Lücken der kommenden KI-Welle.

Dieses Whitepaper beschreibt den Ansatz der **Trust Agents:** KI-Agentensysteme mit systemisch integrierter Governance-, Kontroll- und Compliance-Fähigkeit. Ziel ist es, die Grundprinzipien von „Trustworthy AI“ von einem abstrakten Prinzipienkatalog in ein operativ nutzbares, messbares und auditierbares Steuerungsmodell zu überführen.

Im Zentrum steht die These, dass Vertrauen künftig nicht nur ein ethischer oder regulatorischer Faktor sein wird, sondern eine wirtschaftliche Voraussetzung für die produktive Skalierung autonomer Systeme. Unternehmen werden AI-Agenten nur dann nachhaltig einsetzen können, wenn diese:

- nachvollziehbar handeln,
- kontrollierbar bleiben,

- regulatorische Anforderungen einhalten,
- und ihre Entscheidungen auditierbar dokumentieren.

Trust wird damit zur Betriebserlaubnis autonomer KI-Systeme.

Organisationen, die frühzeitig belastbare Governance- und Trust-Mechanismen etablieren, schaffen nicht nur regulatorische Sicherheit, sondern einen strategischen Wettbewerbsvorteil: Sie werden KI-Agenten schneller, breiter und wirtschaftlich erfolgreicher produktiv einsetzen können als ihre Wettbewerber.

2. Der neue Handlungsdruck: Warum KI-Agenten eine neue Form der Governance erfordern

Die aktuelle Diskussion rund um Generative KI konzentrierte sich in den vergangenen Jahren primär auf Modelle, Produktivität und technische Leistungsfähigkeit. Mit KI-Agenten verschiebt sich diese Perspektive grundlegend. Unternehmen stehen nicht mehr nur vor intelligenten Assistenzsystemen, sondern vor zunehmend autonomen digitalen Akteuren.

Diese Systeme treffen eigenständig Entscheidungen, priorisieren Aufgaben, orchestrieren Prozesse, kommunizieren mit anderen Systemen und agieren teilweise ohne unmittelbare menschliche Interaktion. Dadurch verändert sich nicht nur die technologische Architektur von Unternehmen, sondern auch die Art, wie Verantwortung, Kontrolle und Governance gedacht werden müssen.

Gerade in regulierten Branchen ist diese Entwicklung von besonderer Relevanz. Versicherungen, Banken oder Gesundheitssysteme operieren in hochsensiblen Entscheidungsräumen, in denen Nachvollziehbarkeit, Konsistenz und regelkonformes Verhalten geschäftskritisch sind. Bereits heute stellen sich in ersten Pilotprojekten konkrete Fragen – z.B:

- Darf ein Agent eigenständig Kulanzentscheidungen im Leistungsfall treffen?
- Wie wird sichergestellt, dass vergleichbare Fälle konsistent behandelt werden?
- Wie lassen sich Fehlentscheidungen oder nicht-konformes Verhalten verhindern?
- Wann muss ein Mensch eingreifen?
- Und wie kann gegenüber Aufsicht, Revision oder Kunden nachvollziehbar dokumentiert werden, warum ein Agent eine bestimmte Entscheidung getroffen hat?

Klassische Governance-Ansätze reichen dafür nur bedingt aus. Bestehende KI-Governance konzentriert sich häufig auf Richtlinien, Dokumentation oder Modellbewertung – jedoch nicht auf dynamische, autonome Agentensysteme mit operativer Entscheidungsfähigkeit. Die Folge ist eine zunehmende Governance-Lücke zwischen technischer Innovationsgeschwindigkeit und organisatorischer Kontrollfähigkeit.

Parallel dazu steigt der regulatorische Druck erheblich. Mit dem EU AI Act, neuen Anforderungen von EIOPA, BaFin, EBA und weiteren nationalen Aufsichtsbehörden entstehen erstmals konkrete Erwartungen an Transparenz, Nachvollziehbarkeit, Risikoüberwachung und menschliche Kontrolle autonomer AI-Systeme. Audit- und Nachweisfähigkeit entwickeln sich damit von einer optionalen Governance-Maßnahme zu einer geschäftlichen Grundvoraussetzung. Auch die NATO hat hier bereits einen Rahmen zum Einsatz von KI vorgelegt.

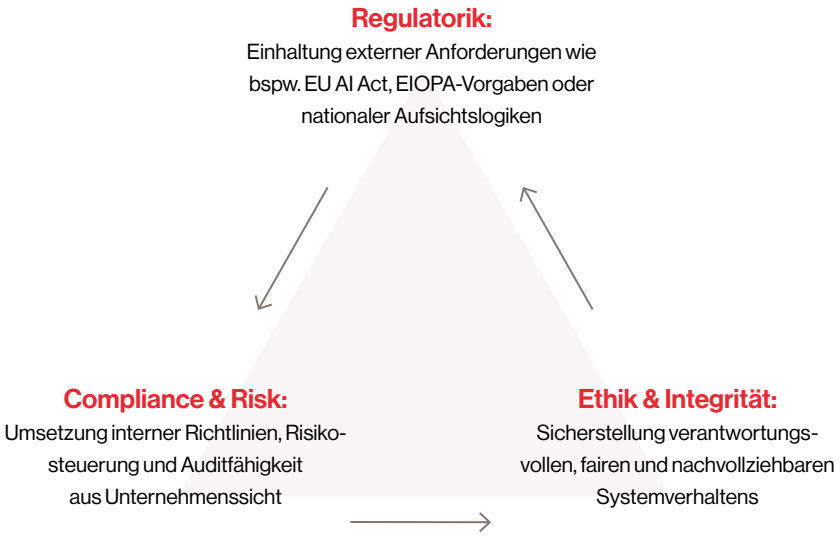
Für Unternehmen entsteht daraus ein strategischer Imperativ:
Die Skalierung autonomer KI-Systeme wird künftig nicht primär an technologischer Leistungsfähigkeit scheitern, sondern an mangelnder Governance, fehlender Kontrollierbarkeit und unzureichender Vertrauensfähigkeit. Genau an diesem Punkt setzt das Konzept der Trust Agents an.

3. Trust Agents: Governed Autonomy als Zielbild

Mit dem Konzept der Trust Agents wird ein neues Zielbild für den Einsatz von KI-Agenten in regulierten Organisationen beschrieben. Im Kern geht es nicht mehr nur um leistungsfähige oder automatisierte Systeme, sondern um **kontrollierbare, nachvollziehbare und regulatorisch anschlussfähige autonome Einheiten,** die sich innerhalb definierter Leitplanken bewegen.

Ein Trust Agent ist damit ein KI-System, das nicht nur Aufgaben ausführt, sondern innerhalb eines klar definierten Governance-Rahmens operiert. Entscheidungen werden nicht nur technisch erzeugt, sondern systematisch begrenzt, überwacht und nachvollziehbar gemacht. Autonomie wird dabei nicht reduziert, sondern gezielt **durch Governance ermöglicht.**

Dieses Zielbild basiert auf drei integrierten Perspektiven, die in regulierten Industrien untrennbar miteinander verbunden sind:



Trust Agents operationalisieren diese Perspektiven in einem konsistenten Steuerungsmodell, das KI-Verhalten nicht nur beschreibt, sondern aktiv formt. Zentrale Prinzipien sind dabei:

- **Governed Autonomy:**
  Definierte Handlungsräume statt unkontrollierter Freiheit
- **Compliance by Design:**
  Regulatorische Anforderungen sind systemisch eingebettet
- **Auditability by Design:**
  Jede Entscheidung ist nachvollziehbar und prüfbar
- **Human Oversight:**
  Gezielte menschliche Eingriffsmöglichkeiten in kritischen Situationen – so viel wie nötig, so wenig wie möglich
- **Quantified Trust:**
  Vertrauensfähigkeit wird messbar und vergleichbar

Damit entsteht ein Rahmen, der KI-Agenten erstmals systematisch steuerbar macht – nicht über einzelne Regeln oder Policies, sondern über ein konsistentes, messbares Governance-Modell.

Die zentrale Innovation liegt in der Operationalisierung von Vertrauen. Trust wird nicht länger als abstraktes Prinzip verstanden, sondern als **strukturierte Kombination messbarer Dimensionen,** die das Verhalten von Agenten beschreiben und begrenzen.

Dazu zählen insbesondere:
Transparenz, Erklärbarkeit, Konsistenz und Reproduzierbarkeit, Kontrollierbarkeit, Fairness, Risikosensitivität, Auditierbarkeit und Robustheit sowie strategic alignment.

In ihrer Gesamtheit bilden diese Dimensionen die Grundlage für ein quantifiziertes Trust-Modell, das AI-Agenten in unterschiedliche Vertrauens- und Kontrollstufen einordnet. Dadurch entstehen erstmals vergleichbare und steuerbare Systeme, die sowohl operativ einsetzbar als auch regulatorisch anschlussfähig sind.

4. Business Value:
Was ist der Wert von Trust Agents?

Der Einsatz von Trust Agents adressiert nicht nur eine regulatorische Notwendigkeit, sondern eröffnet einen unmittelbaren wirtschaftlichen Hebel für Organisationen in regulierten Branchen.

Im Zentrum steht dabei die Fähigkeit, KI-Agenten **schneller, sicherer und in größerem Umfang produktiv zu setzen.** Während viele Organisationen heute noch in Pilot- oder Experimentierphasen verharren, wird Governance zur entscheidenden Voraussetzung für Skalierung.

Der konkrete Business Value zeigt sich in mehreren Dimensionen:

Beschleunigung der KI-Produktivsetzung	**Schnellere regulatorische Konformität**
Reduktion von Freigabe- und Abstimmungszyklen durch klare Governance-Strukturen	Vereinfachung von Prüf- und Freigabeprozessen gegenüber Aufsicht und Revision
Reduktion operativer Risiken	**Skalierbarkeit autonomer Prozesse**
Geringere Fehleranfälligkeit und kontrollierte Entscheidungsräume	Übertragbarkeit von Agenten auf weitere Use Cases ohne Governance-Neuentwicklung
Senkung von Compliance-Kosten	**Erhöhte Stakeholder-Akzeptanz**
Weniger manuelle Prüfaufwände durch integrierte Auditfähigkeit	Höhere Vertrauensfähigkeit gegenüber Kunden, Aufsicht und internen Kontrollinstanzen

Strategisch verändert sich damit die Rolle von Governance grundlegend. Sie wird nicht länger als nachgelagerte Kontrollfunktion verstanden, sondern als **Enabler für skalierbare Autonomie.** Organisationen, die Trust systematisch operationalisieren, schaffen die Voraussetzung für eine neue Form der Unternehmenssteuerung: die „Governed Enterprise Autonomy".

In diesem Modell wird Vertrauen selbst zu einem Wettbewerbsfaktor. Unternehmen, die KI-Agenten nicht nur leistungsfähig, sondern auch kontrollierbar und auditierbar gestalten, können diese schneller und breiter einsetzen als Wettbewerber mit rein technischer oder experimenteller AI-Strategie.

Parallel dazu verdichtet sich der regulatorische Kontext. Mit dem EU AI Act, neuen Leitlinien von EIOPA, BaFin EBA und weiteren Aufsichtsbehörden steigt die Erwartung an Nachweisbarkeit, Transparenz und Risikosteuerung kontinuierlich an. Auditierbarkeit entwickelt sich damit zu einer strukturellen Marktzugangsvoraussetzung für den produktiven Einsatz autonomer Systeme.

Trust Agents adressieren genau diese Entwicklung: Sie verbinden wirtschaftliche Skalierbarkeit mit regulatorischer Sicherheit und schaffen damit die Grundlage für den produktiven Einsatz autonomer AI in hochregulierten Industrien.

5. Operating Model: So gelingt die strategische Steuerung von Agenten

Die Einführung von Trust Agents erfordert nicht nur technologische Anpassungen, sondern ein neues **operatives Steuerungsmodell für autonome KI-Systeme.** Klassische IT-, Risk- und Compliance-Strukturen sind dabei nicht ausreichend, da sie primär auf statische Systeme und dokumentenbasierte Kontrolle ausgelegt sind.

Das **Trust Agent Operating Model** etabliert daher eine mehrschichtige Governance-Architektur, die AI-Agenten entlang ihres gesamten Lebenszyklus steuert:

- **Governance Guadrails:**
  Definition von Regeln, Entscheidungsgrenzen und operativen Leitplanken für Agenten
- **Risk- & Ethics Layer:**
  Kontinuierliche Bewertung von Risiken, Sensitivitäten und potenziellen Fehlverhalten
- **Monitoring Komponenten:**
  Laufende Überwachung von Entscheidungen, Verhalten und Performance der Agenten
- **Human Escalation Pathways:**
  Definierte Eingriffs- und Kontrollpunkte für menschliche Entscheidungsträger
- **Audit Trails:**
  Vollständige Dokumentation und Nachvollziehbarkeit aller relevanten Entscheidungen

Dieses Modell ermöglicht erstmals eine systematische Integration von AI-Agenten in bestehende Unternehmenssteuerung, ohne Kontrollverlust zu riskieren. Entscheidend ist dabei die Verschiebung von punktueller Kontrolle hin zu **kontinuierlicher, systemischer Governance.**

Auf organisatorischer Ebene entstehen daraus neue Rollen und Verantwortlichkeiten. Unternehmen benötigen künftig explizite Funktionen für AI-Überwachung, Governance und Lifecycle-Steuerung autonomer Systeme. Beispiele sind AI Oversight Boards, Agent Lifecycle Management sowie spezialisierte Governance- und Risk-Funktionen für AI-Systeme.

Damit wird AI nicht mehr als rein technologische Komponente betrachtet, sondern als **dauerhaft zu steuernder Organisationsbestandteil** mit klaren Kontroll- und Verantwortungsstrukturen.

6. Was nun zu tun ist:
Handlungsempfehlungen für Executives

Die erfolgreiche Einführung von Trust Agents erfordert ein schrittweises, strategisch gesteuertes Vorgehen, das technologische Entwicklung, Governance-Aufbau und regulatorische Anschlussfähigkeit integriert.

Kurzfristig

Im ersten Schritt sollten Organisationen die relevanten KI-Agenten-Use-Cases identifizieren, die bereits heute oder in naher Zukunft in kritischen Entscheidungsprozessen eingesetzt werden. Parallel dazu sind erste Governance-Anforderungen zu definieren sowie initiale Kontroll- und Eskalationsmechanismen zu pilotieren.

Mittelfristig

Im nächsten Schritt steht der Aufbau eines konsistenten Trust Agents Frameworks im Mittelpunkt. Dieses bildet die Grundlage für einheitliche Bewertungs- und Steuerungslogiken über verschiedene KI-Agenten hinweg. Gleichzeitig müssen Audit- und Nachweisfähigkeiten strukturell aufgebaut und in bestehende Compliance- und Risk-Prozesse integriert werden. Ziel ist die Operationalisierung von KI Governance als wiederkehrender Standardprozess.

Langfristig

Langfristig entwickeln sich Trust Agents zu einem festen, strategischen Bestandteil der Unternehmensarchitektur. Governance wird dabei nicht mehr als Zusatzfunktion verstanden, sondern als Kernfähigkeit der Organisation. Zielbild ist eine vollständig **Governed Enterprise Autonomy,** in der KI-Agenten skalierbar, kontrollierbar und regulatorisch abgesichert operieren.

Damit wird Trust nicht nur zu einem Kontrollmechanismus, sondern zu einer strategischen Fähigkeit für den nachhaltigen Einsatz autonomer Systeme in regulierten Industrien.

7. Fazit & strategischer Ausblick:
Trust als neue Systemlogik

Die Einführung von KI-Agenten markiert keinen inkrementellen Technologieschritt, sondern einen strukturellen Wandel in der Art, wie Organisationen Entscheidungen automatisieren und operative Verantwortung gestalten. Erstmals entstehen Systeme, die nicht nur unterstützen, sondern eigenständig handeln – und damit unmittelbar in geschäftskritische Prozesse eingreifen.

In diesem Kontext verschiebt sich der Erfolgsfaktor für den Einsatz von AI fundamental. Nicht mehr die Leistungsfähigkeit einzelner Modelle oder die Geschwindigkeit der Implementierung entscheidet über Wertschöpfung, sondern die Fähigkeit, diese Systeme **kontrollierbar, nachvollziehbar und vertrauenswürdig zu betreiben.**

Gerade in regulierten Branchen wird damit eine neue Differenzierung sichtbar: Während technologische Fähigkeiten zunehmend commoditisiert werden, wird Governance zur eigentlichen Skalierungsbarriere – und zugleich zur entscheidenden Enabler-Kompetenz.

Organisationen, die frühzeitig in strukturierte Trust- und Governance-Mechanismen investieren, schaffen damit nicht nur regulatorische Sicherheit, sondern die Grundlage für eine neue Form der Unternehmenssteuerung: **skalierbare, aber kontrollierte Autonomie von KI-Systemen.**

Die Konsequenz ist klar: Die nächste Phase der KI-Transformation wird nicht durch die leistungsfähigsten Agenten dominiert, sondern durch diejenigen Organisationen, die ihre Agenten am besten steuern, absichern und in ihre bestehenden Risiko- und Kontrollsysteme integrieren können.

Damit entsteht ein neuer strategischer Imperativ für regulierte Industrien:
KI-Agenten sind keine reine Technologiefrage mehr – sie sind eine Frage von Betriebserlaubnis, Governance-Reife und institutionellem Vertrauen.

8. Quellen & Further Reading:

Ethics Guidelines for trustworthy AI der EU-Kommission:
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

AI Governance Principles der EIOPA:
https://www.eiopa.europa.eu/system/files/2021-06/eiopa-ai-governance-principles-june-2021.pdf

EU AI Act:
https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

NATO: Responsible AI use in Defense:
https://regulations.ai/regulations/RAI-X6-GO-RESPONS-2024

RE-centric Recommendations for the Development of Trustworthy(er) Autonomous Systems:
https://arxiv.org/pdf/2306.01774

Unsere Autoren:

Dr. Markus Knappitsch
Executive Manager
Comma Soft AG
markus.knappitsch@comma-soft.com

Comma Soft AG
Pützchens Chaussee 202 – 204a
D-53229 Bonn
comma-soft.com

Prof. Dr. Markus Gabriel
Founder
deep-in GmbH
markus.gabriel@deep-in.ai

Deep-IN GmbH
Thomas-Mann-Str. 36
D-53111 Bonn
deep-in.ai