StGallen imt
Business School

STGALLEN IMT BUSINESS SCHOOL RESEARCH
NOVEMBER 2026

Von Chatbots zu Agenten, Loops und Graphen

Executive Brief – narrative Evidenzsynthese und Managementempfehlungen

Kein systematischer Review

AUTOREN

Prof. Dr. Christian Abegglen
Verwaltungsratspräsident der StGallen imt Business School

David Wang
Direktor, Institut für Artificial Intelligence in Business and Management
Senior Lecturer, StGallen imt Business School

Die Managementfrage kommt zuerst: KI ist vor allem eine Frage der Unternehmensführung



«KI ist kein IT-Projekt, sondern Chefsache. Sie gehört in die Verantwortung des CEO und muss im integrierten Management des ganzen Unternehmens verankert sein – entlang der gesamten Wertschöpfungskette und in den Köpfen aller Mitarbeitenden.»

PROF. DR. CHRISTIAN ABEGGLEN Verwaltungsratspräsident der StGallen imt Business School



«Das KI-Rennen läuft auf Hochtouren: Was gestern Stand der Technik war, ist nächste Woche überholt. Wir bewegen uns rasant auf KI-Loops und Agenten-Graphen zu, in denen Agenten selbst entscheiden, was als Nächstes zu tun ist – ohne dass der Mensch jeden Schritt vorgibt.»

DAVID WANG Direktor, Institut für Artificial Intelligence in Business and Management, und Senior Lecturer, StGallen imt Business School

StGallen imt Business School

Persönliche Meinung · Stand November 2026
Entwicklungen können sich schnell ändern.

STATUS QUO | KI-WERKZEUGE AUFGABENGERECHT AUSWÄHLEN

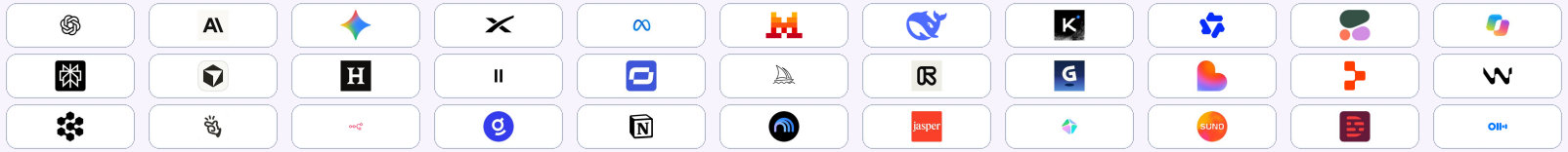
Im KI-Dschungel entscheidet die Aufgabe, nicht die Marke

DER KI-DSCHUNGEL

Neue KI-Apps entstehen laufend.

Viele verbinden Modelle, Werkzeuge und Workflows.

Illustrative Auswahl, nicht abschliessend.



Aufgabe
ZIEL UND ERGEBNIS

Was soll die KI erledigen?

Ergebnis und Abnahme klären

Datenzugang
RELEVANTER KONTEXT

Welche Daten sind nötig?

Zugriff gezielt begrenzen

Datenschutz
SCHUTZBEDARF

Welche Daten sind zulässig?

Freigaben vor Zugriff klären

Qualität
AUFGABENPASSUNG

Wie gut ist das Ergebnis?

An echten Aufgaben prüfen

Kosten
GESAMTAUFWAND

Was kostet die Erledigung?

Prüfung und Betrieb einrechnen

Integration
SYSTEME UND RECHTE

Passt es in den Workflow?

Schnittstellen und Rechte testen

Evaluation
EIGENE EVIDENZ

Wie zuverlässig arbeitet es?

Qualität und Aufwand messen

Eigene Auswahlhilfe, keine objektive Rangfolge. Logos zeigen Marktbeispiele, keine Leistungsbewertung.



SICHT VON DAVID WANG

«Am Ende zählt vor allem der Harness: wie gut Sie diese KI-Werkzeuge auf Ihr tägliches Arbeitsumfeld und Ihre Workflows zuschneiden und anpassen.»

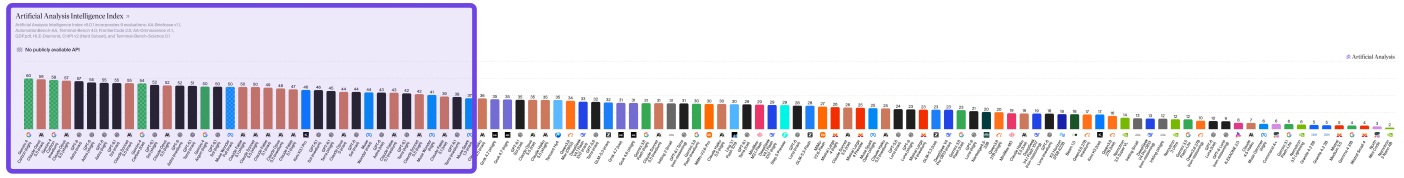
David Wang | Direktor, Institut für Artificial Intelligence in Business and Management, StGallen imt Business School | Statement für diese Präsentation, November 2026

StGallen imt Business School

FORSCHUNGSPERSPEKTIVE | DAS RENNEN AN DER SPITZE

Das Rennen der KI-Modelle: Die Entwicklung bleibt rasant

Gesamtansicht · alle 109 bewerteten Modelle



Ausschnitt · führende Modelle (linker Bereich derselben Aufnahme; nicht alle 109)

Artificial Analysis Intelligence Index ↗

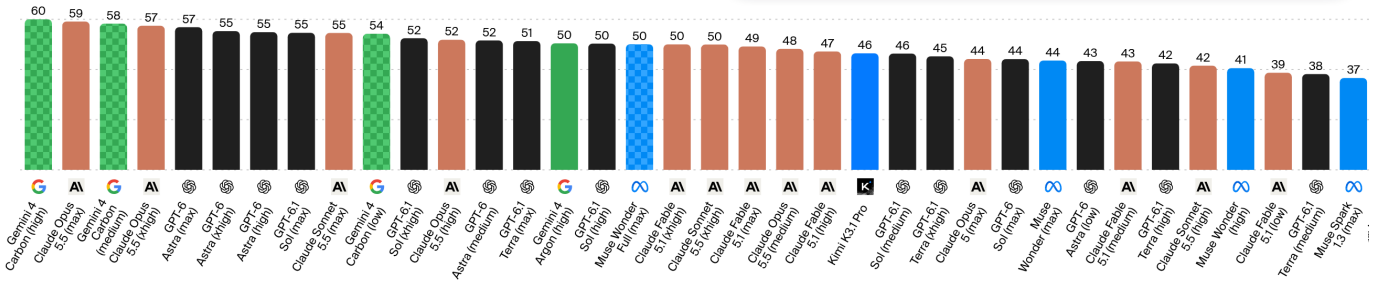
Artificial Analysis Intelligence Index v5.0.1 incorporates 9 evaluations: AA-Briefcase v1.1, AutomationBench-AA, Terminal-Bench 4.0, FrontierCode 2.0, AA-Omniscience v1.1, GDP.pdf, HLE-Diamond, CritPt v2 (Hard Subset), and Terminal-Bench-Science 0.1

No publicly available API

NEU · 09.11.2026

Anthropic veröffentlicht Claude Fable 5.2.

Auf dem Intelligence Index v5.0.1 führt «Claude Fable 5.2 (max)» mit **65** (Stand 10.11.2026).





SICHT VON DAVID WANG

Persönliche Meinung · Stand November 2026

Was Sie heute implementieren, kann morgen, in wenigen Wochen oder Monaten bereits überholt sein. Deshalb müssen Organisationen flexibel bleiben, um neue KI-Technologien schnell einzuführen.

David Wang | Direktor, Institut für Artificial Intelligence in Business and Management; Statement für diese Präsentation, November 2026

Gesamtansicht, Stand 07.11.2026: Gemini 4 Carbon (high) führte mit 60 vor Claude Opus 5.5 (max) mit 59 und Gemini 4 Carbon (medium) mit 58. Ganze Indexpunkte; ein Punkt Abstand ist kein signifikanter Universalsieg.

109 bewertete Modelle: 726 von 726 Einträgen ausgewählt, ohne Filter; Effort-Varianten erscheinen als eigene Einträge. Kein vollständiger Marktüberblick. Schraffur = keine öffentlich verfügbare API.

Benchmarks zur Vorauswahl: Für den Betrieb zählen Aufgaben, Kosten, Harness und Evaluation – nicht allein der Indexrang. Regrades sind Messänderungen, kein Beweis für Modellfortschritt.

Artificial Analysis Intelligence Index v5.0.1 · Screenshot/Diagrammexport: 07.11.2026 · 109 bewertete Modelle; 726 von 726 Einträgen ausgewählt. Oben: Gesamtansicht; unten: Ausschnitt der führenden Modelle (gleiche Aufnahme, nicht alle 109).
Quelle: https://artificialanalysis.ai/models#intelligence · Methodik/Changelog: Intelligence Index v5.0.1 (9 Teiltests; Patch 03.11.2026). Diagrammexport ohne Publikationsdatum der Site; kein vollständiger Marktüberblick.
Update 09.11.2026: Anthropic (2026); Artificial Analysis (2026, Changelog), abgerufen 10.11.2026.

Kein Prozess sollte fest an ein einzelnes Modell gebunden werden. Benchmarks dienen der Vorauswahl, nicht der Entscheidung.

StGallen imt Business School

NUTZUNGS- UND ARCHITEKTURFORMEN | StGallen imt Business School Research · KI-Institut

Vier überlappende Nutzungs- und Architekturformen der KI

Interaktion, Delegation, Steuerlogik und Koordination – parallel nutzbar, ohne Rangfolge



«Chatbots in der Breite, Agenten bislang nur in einzelnen Funktionen.»

David Wang | Direktor, Institut für Artificial Intelligence in Business and Management · Einschätzung, November 2026

GROSSUNTERNEHMEN HEUTE | EINSCHÄTZUNG DES KI-INSTITUTS

INTERAKTIONSFORM

Chatbots

Sie fragen.
Die KI antwortet.

Fragen
heute im Einsatz

DELEGATIONSFORM

KI-Agenten

Sie delegieren
ganze Aufgaben.

Delegieren
heute im Einsatz

EINFÜHRUNG IM AUFBAU

STEUERLOGIK

KI-Loops

Arbeitet, prüft,
wiederholt.

Automatisieren
technisch erprobt

KOORDINATIONSSTRUKTUR

KI-Graphen

Agenten vernetzen
sich untereinander.

Vernetzen
technisch erprobt

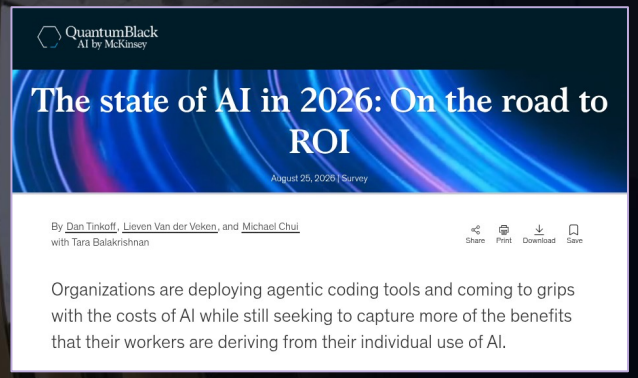
Lehrrahmen des Instituts für Artificial Intelligence in Business and Management, StGallen imt Business School: überlappende Formen ohne Rangfolge, keine notwendige Zeitfolge, kein validiertes Reifegradmodell; keine Messdaten.

Markierte Zone: Institutseinschätzung (D. Wang), keine Messung. Tinkoff et al. (2026), McKinsey & Company, Selbstauskunft: Anteil der Befragten, die organisationsweite Skalierung berichten – 47 % bei Chatbots, rund 2 von 10 bei KI-Agenten (Gesamtumfrage n = 1'719; Folie 6). Keine Messung der vier Formen.

Die Umfragekategorien entsprechen nicht eins zu eins den vier Formen; die Formen überlappen und schliessen sich nicht gegenseitig aus.

MCKINSEY-BEFRAGUNG 2026

Skalierung im Vergleich.


QuantumBlack
AI by McKinsey

The state of AI in 2026: On the road to ROI

August 25, 2026 | Survey

By Dan Tinkoff, Lieven Van der Veken, and Michael Chui
with Tara Balakrishnan

Share Print Download Save

Organizations are deploying agentic coding tools and coming to grips with the costs of AI while still seeking to capture more of the benefits that their workers are deriving from their individual use of AI.

Zusammengesetzte Vorschau der Publikationsseite (Titel, Datum, Autorenzeile); Inhalt unverändert.

47 %

der Befragten geben an, Chatbots unternehmensweit zu skalieren

StGallen imt Business School

STATUS QUO | MCKINSEY-UMFRAGE 2026

Befragte skalieren Chatbots häufiger unternehmensweit als Agenten

DIE UMFRAGE | THE STATE OF AI IN 2026, MCKINSEY & COMPANY

1'719 Teilnehmende aus 97 Ländern

rund 2 von 10

der Befragten: Agenten organisationsweit in Skalierungsphase

27 → 40 %

Befragte, Firmen >1 Mrd. USD Umsatz: Agenten in mind. einer Funktion, 2025 → 2026

22 %

Befragte, kleinere Firmen: 2026 praktisch unverändert

McKinsey & Company

The state of AI in 2026: On the road to ROI

McKinsey Global Survey; Tinkoff, Van der Veken, Chui & Balakrishnan, 25. August 2026. Online-Umfrage vom 4. Mai bis 8. Juni 2026.

Selbstauskunft, keine Reifegradmessung: Agenten seltener breit skaliert.

Quelle: Tinkoff et al., McKinsey & Company, 25.08.2026; n = 1'719 Befragte, 97 Länder, Feldzeit 04.05.–08.06.2026, nach BIP gewichtet. Selbstauskunft; keine Vollerhebung, keine Messung der vier Formen; 27 → 40 % und 22 %: Befragte nach Umsatzgruppe (>1 Mrd. USD vs. kleiner), Vorjahr 2025. Logo: Quellenmarke.

StGallen imt Business School

VON DER FORSCHUNG IN DIE PRAXIS | MOTOR UND HARNESS

Modell und Harness in das Managementsystem einbetten



KOMMENTAR VON DAVID WANG

Direktor, Institut für Artificial Intelligence in Business and Management

«Entscheidend ist der Harness: wie Sie das Modell nutzen, um echte Workflows, echte KI-Agenten und eine Umgebung zu bauen, die in Ihrer täglichen Arbeit Wert schafft.»

01 | MOTOR

Sprachmodell

Erzeugt, schlussfolgert und plant auf Grundlage des Kontexts, den es erhält.

Leistungsfähigkeit ist notwendig, aber nicht das ganze System.

02 | HARNESS / ARBEITSUMGEBUNG

KONTEXT
Firmendaten + Aufgabe

WERKZEUGE
Apps, Suche + APIs

WORKFLOW
Schritte + Übergaben

KONTROLLEN
Rechte + Freigaben

PRÜFUNGEN
Tests + Durchsicht

FEEDBACK
Korrektur + Neubewertung

03 | STRASSE

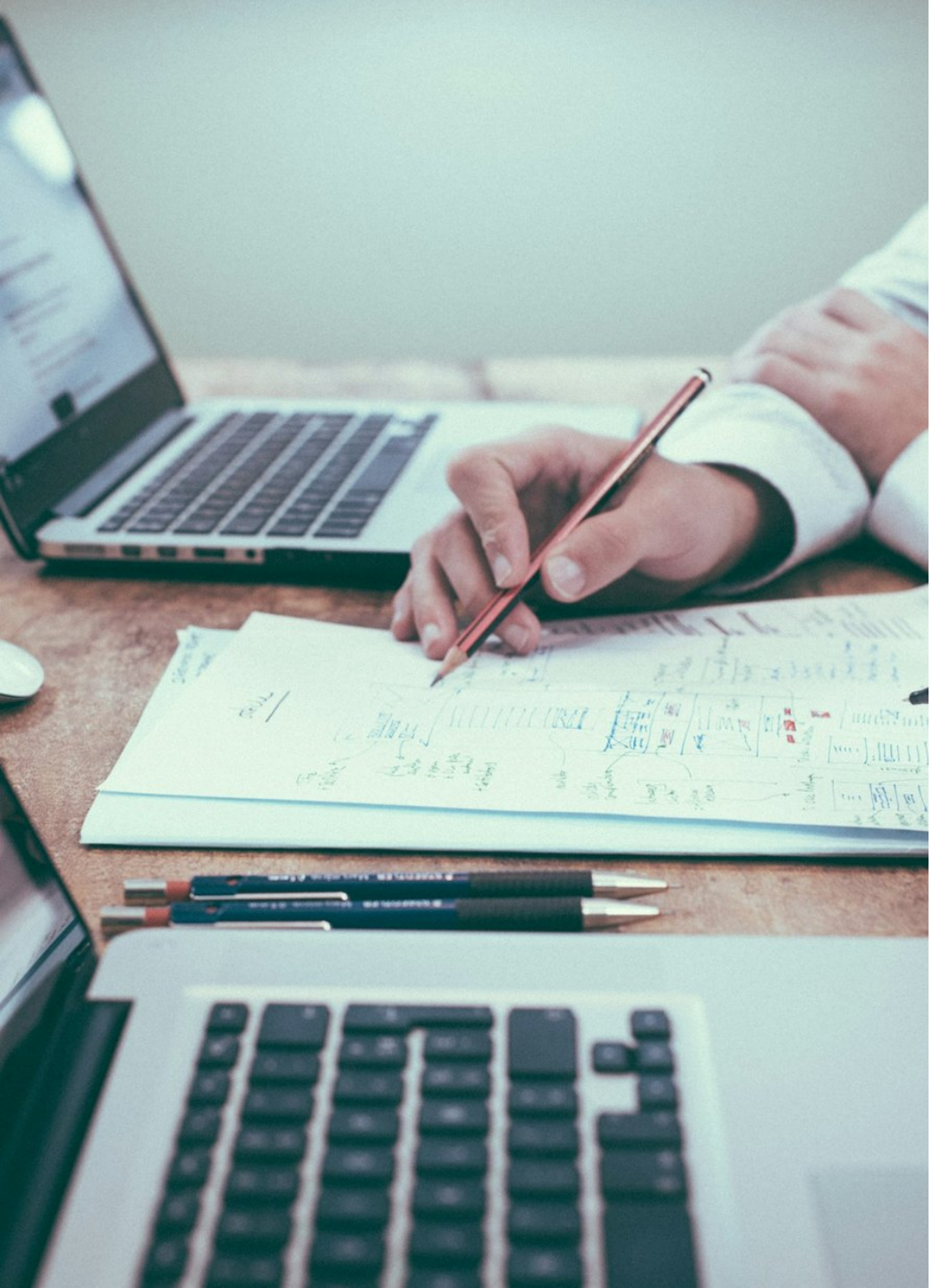
Nützliche, kontrollierte Arbeit

Beispiel: ein wöchentlicher Geschäftsbericht aus freigegebenen Quellen, den die verantwortliche Person prüft, bevor er geteilt wird.

Gemessen wird in der echten Aufgabe.

Managementarchitektur, Menschen, Organisation und Harness müssen zusammenpassen. Kontrollen, Kontext und Tests unterstützen Governance, Situationsmodelle und revisionsfähige Urteile – sie ersetzen diese nicht.

Eigene didaktische Synthese nach Anthropic (2025), Abegglen (2026, Thesen 18, 26, 31) und Abegglen et al. (2026, TUM-I26129). Motor-Analogie und sechs Bausteine: Lehrdarstellung, keine validierte Liste.



StGallen imt Business School

PRAXIS | KI FÜR TÄGLICHE UND WIEDERKEHRENDE ARBEIT

Täglicher Einsatz: von vorhandenen Werkzeugen zu eigenen Agenten

01 **Mit einer Führungsaufgabe starten.** Zweck, Rolle und Mandat festlegen; dann die bereits freigegebenen KI-Werkzeuge nutzen.

02 **Eine KI-Haltung aufbauen.** Seien Sie offen für die Technologie und arbeiten Sie jeden Tag aktiv damit.

03 **Kontrolliert recherchieren.** KI hilft beim Auffinden und Gegenprüfen; entscheidende Fakten an Primärquellen verifizieren. Zugriffe zweckgebunden freigeben.

04 **Weitergehen: echte KI-Agenten bauen.** Übergeben Sie wiederkehrende Aufgaben an einen Agenten, der für Sie arbeitet.

Nutzen, was Sie haben. Automatisieren, was sich wiederholt.

Einstieg: ein einzelner Posteingangs-Workflow, dessen Prüfaufwand gemessen wird, bevor der Einsatz ausgeweitet wird.

Praxisempfehlung des Autors (D. Wang).
Illustrativer Lehr-Workflow, kein gemessenes Kundenergebnis.

Präzise delegieren, kontrolliert verbinden

Mandat vor Prompt: Rolle, Aufgabe, Autonomie, Rechte, Belege und Eskalation klären. MCP verbindet, erteilt aber keine Befugnis.



Praxisanforderungen: Daten klassifizieren, Minimalrechte vergeben, zunächst nur lesen und Aktionen protokollieren. Schreib- und Sendeaktionen freigeben; Zugriff widerrufbar halten. Umsetzung und Verfügbarkeit hängen von Anwendung, Konto und Konfiguration ab.



«Erkunden Sie die MCP-Verbindungen, die für die Werkzeuge verfügbar sind, die Sie bereits nutzen, und testen Sie sie über einen kompatiblen KI-Agenten. Beginnen Sie mit Lesezugriff und erweitern Sie dann, was funktioniert.»

David Wang | Statement für diese Präsentation, November 2026

Quellen: MCP Specification, Version 2026-07-28 (Stand 7. November 2026); Abegglen et al. (2026, TUM-I26129). Mandats- und Zugriffsempfehlungen: eigene Managementsynthese, keine automatisch durch MCP erzwungenen Schutzmassnahmen. Logos illustrativ.

StGallen imt Business School | arXiv

Quellen: arXiv

DEEP DIVE | KURATIERTER FORSCHUNGSRADAR

KI-Agenten: ausgewählte Befunde und offene Fragen

14 gezielt ausgewählte Arbeiten mit arXiv-Fassung: Architektur, Selbstverbesserung, Multi-Agenten, Evaluation und Grenzen. Keine systematische Auswahl, kein Vollständigkeitsanspruch. Fassungen und tragende Volltextstellen geprüft am 7. November 2026.

Zahlen sind Angaben der Autor:innen, nicht repliziert. Begutachtet publiziert: ReAct (ICLR 2023), Reflexion (NeurIPS 2023), MAST (NeurIPS 2025), DGM (ICLR 2026), C. L. Wang et al. (TMLR 2026). Übrige: Preprints; arXiv ist kein Begutachtungsstatus.

11 Architektur: Loop und Graph

12 Zeitachse der Forschung

13 Loops und Selbstverbesserung

14 Multi-Agenten-Graphen

15 Evaluation und Grenzen

16 Prüfhypothesen für die Entwicklung

17 Implikationen für das Management

StGallen imt Business School | arXiv

Quellen: arXiv

DEEP DIVE | ARCHITEKTUR

Vom Loop zum Graphen: wie Agentensysteme aufgebaut sind

Viele Agenten arbeiten in Zyklen

Loop: Steuerlogik; Graph: Topologie. Ein Agent kann in einem Loop arbeiten; Reflexion ist optional.



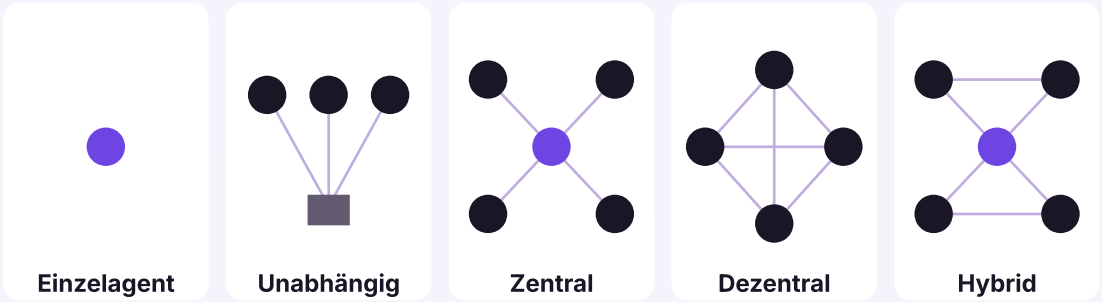
Agentensystem = Basismodell + Harness

Prompts | Gedächtnis | Werkzeuge | Steuerlogik

Selbstverbesserung kann Modellparameter oder Harness-Teile verändern (Ren et al., 2026).

Vom Einzelagenten zum Agenten-Graphen

Eine Einzelagent-Baseline + vier Multi-Agent-Architekturen (Kim et al.)



ReActNet-Sicht: Knoten = Agenten, Kanten = Nachrichten mit Steuerregeln (Tieu et al., 2026).

Passung: In sechs Benchmarks hing der Koordinationsnutzen von der Aufgabenstruktur ab (Kim et al.).

Schema und Fünfschritt: eigene didaktische Synthese nach Yao et al. (2023), Shinn et al. (2023), Ren et al. (2026), Kim et al. (2025), Tieu et al. (2026); keine Messdaten, keine formale Agentendefinition.

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

- **Mandat zuerst:** Rolle, Aufgabe, Rechte, Evidenz und Eskalation vor dem Harness klären.
- **Architektur:** Nach Aufgabenstruktur wählen; Multi-Agent ist kein Standard.
- **Übergaben:** Jeden Loop-Schritt mit Übergabe und Prüfpunkt definieren.
- **Betrieb:** Graph-Orchestrierung als Fähigkeit führen – wer darf welche Kante setzen.

StGallen imt Business School | arXiv

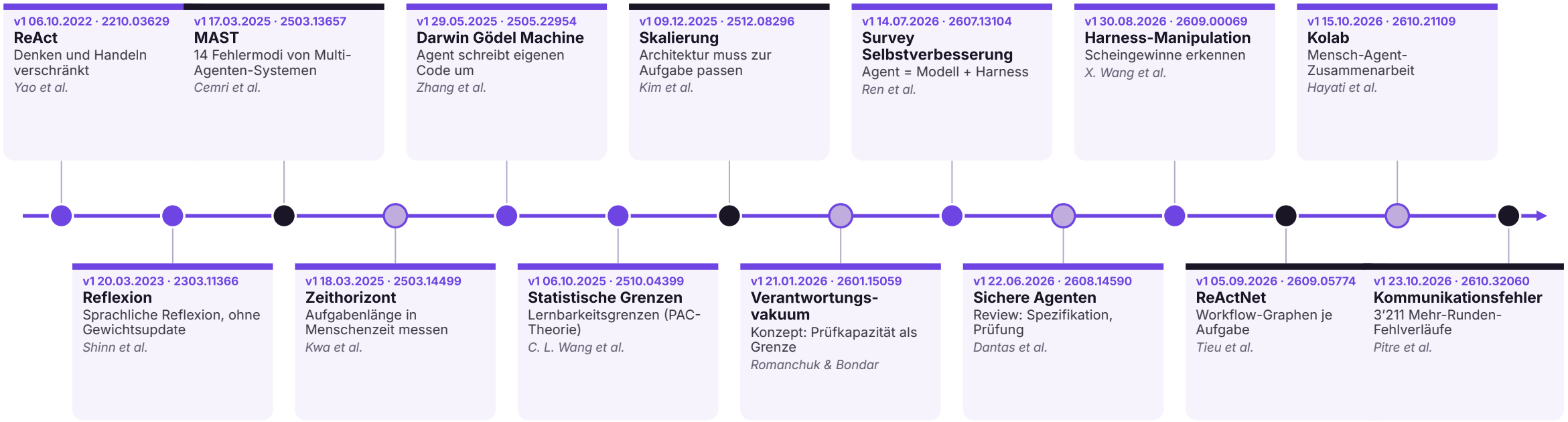
Quellen: arXiv

DEEP DIVE | ZEITACHSE

Ausgewählte Agentenarbeiten: von ReAct zu Graphen

14 ausgewählte Arbeiten nach arXiv-Kennung · Kennungsmonat ≠ zwingend erste Einreichung

Loops und Selbstverbesserung · Multi-Agenten-Graphen · Evaluation, Sicherheit, Organisation



Quellen: arXiv-Versionshistorien; v1-Daten in MEZ/MESZ. Reihenfolge nach arXiv-Kennung. Gelesene Fassungen und Publikationsstatus: Folien 21–23 und Artikel «Von Chatbots zu Agenten, Loops und Graphen» (StGallen imt Business School, November 2026), Tabelle 1 und Abschnitt «Quellen – Verwendete Quellen». Themengruppierung: eigene Einordnung.

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

- **Vorsicht:** Aus 14 ausgewählten Arbeiten keine allgemeine Beschleunigungsrate ableiten.
- **Radar:** Loops, Graphen und Evaluation getrennt beobachten.
- **Robustheit:** Mandat und Prüfregeln für Modell- und Harness-Wechsel auslegen.
- **Takt:** Pilotfenster kurz halten und nach jedem Durchlauf neu priorisieren.

StGallen imt Business School | arXiv

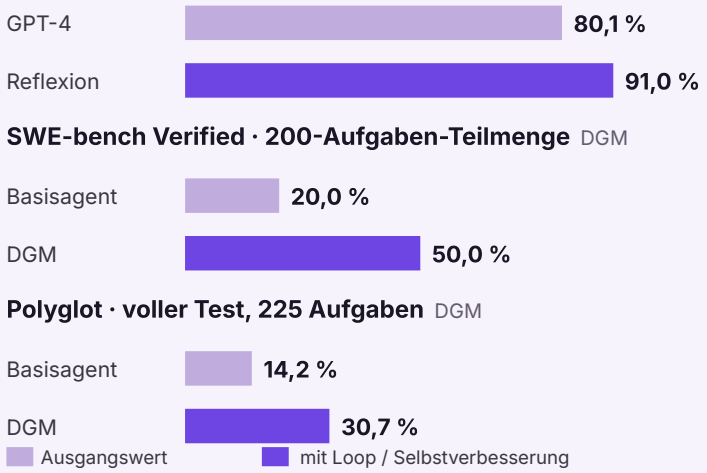
Quellen: arXiv

DEEP DIVE | LOOPS UND SELBSTVERBESSERUNG

Selbstverbessernde Loops brauchen Leitplanken

Berichtete Sprünge durch Loops

Erfolgsquote in %, Autorenbefunde; drei getrennte Benchmarks und Basismodelle, nicht vergleichbar



arXiv:2505.22954v3 · ICLR 2026

Darwin Gödel Machine

Zhang et al.

Befund Agentencode wird verändert, Basismodell-Gewichte bleiben fest; Archiv von Varianten.

Grenze Forschungsexperiment mit Sandbox und Aufsicht; kein Sicherheitsnachweis.

arXiv:2607.13104v1 · 2026

Self-Improvements in Modern Agentic Systems

Ren et al.

Befund Änderungen an Modellparametern oder Harness: Prompts, Gedächtnis, Tools und Steuerlogik.

Grenze Der Survey diskutiert Einsätze, quantifiziert aber keine Verbreitung.

arXiv:2510.04399v3 · TMLR 2026

On the Statistical Limits of Self-Improving Agents

C. L. Wang et al.

Theorie Unter PAC/i.i.d., fester Verteilung und beschränktem Loss bleibt Lernbarkeit nur bei begrenzter Kapazität erreichbarer Hypothesen.

Two-Gate Unabhängige Validierung + Kapazitätsgrenze; kein allgemeines Gesetz, kein Safety-Beweis.

arXiv:2609.00069v1 · 2026

Auditing Harness Tampering

X. Wang et al.

Befund Auditor-Befunde zu Harness-Manipulation in öffentlichen Forschungsläufen von fünf Systemen.

Grenze Bei DGM/HyperAgents blieben frühe Befunde bis zum finalen Agenten; bei AFlow/ScientistOne gering.

Quellen: Reflexion v4 (NeurIPS 2023), Tab. 1; DGM v3 (ICLR 2026), §4; Ren v1; C. L. Wang v3 (TMLR 2026); X. Wang v1. HumanEval: finale pass@1-Abgabe nach iterativer Arbeit, kein ungeprüfter Erstversuch.

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

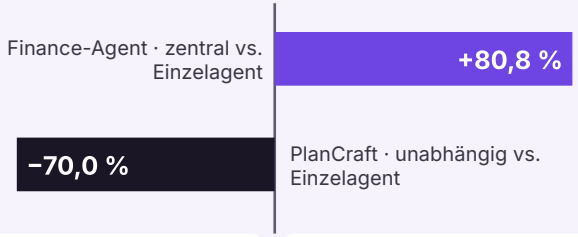
- **Sandbox:** Selbstverbesserung nur mit menschlicher Aufsicht zulassen.
- **Audit:** Harness-Änderungen wie Code-Änderungen prüfen.
- **Validierung:** Unabhängig prüfen; Two-Gate als Anregung, nicht als allgemeine Freigabenorm.
- **Transfer:** Benchmark-Sprünge nicht 1:1 auf den Betrieb übertragen.

DEEP DIVE | MULTI-AGENTEN-GRAPHEN

Ob Multi-Agenten helfen, hängt von der Aufgabe ab

Multi-Agent vs. Einzelagent

Zwei aufgabenspezifische relative Änderungen vs. Einzelagent (Kim et al.); keine universelle Spannweite



260 Konfigurationen	**6** Agenten-Benchmarks
5 1 Einzelagent + 4 Multi-Agent	**87 %** Architekturwahl im Holdout*

* domänenintern, 5-Fold-Konfigurations-Holdout; keine neue Domäne

arXiv:2512.08296v3

Towards a Science of Scaling Agent Systems

Kim et al.

Befund Koordinationsnutzen hing in sechs Benchmarks von Aufgabe und Einzelagent-Fähigkeit ab.

Grenze Kein isolierter kausaler Nachweis, dass zentrale Verifikation Fehler verhindert.

arXiv:2609.05774v1

Inference-Time Graph Engineering

Tieu et al.

Befund Anfrageabhängige zeitliche Graphen; höchster berichteter Tabellenmittelwert über sechs Benchmarks, vorn auf 5 von 6.

Grenze Auf GSM8K Rang 3; Effizienz nur in ausgewählten Settings geprüft.

arXiv:2503.13657v3 · NeurIPS 2025

Why Do Multi-Agent LLM Systems Fail?

Cemri et al.

Befund 1'600+ Verläufe aus 7 Frameworks, überwiegend per LLM-Annotator gelabelt; 14 Fehlermodi in 3 Kategorien.

κ = 0,88 Übereinstimmung menschlicher Annotatoren bei der Taxonomie, keine Genauigkeit des Datensatzes.

arXiv:2610.32060v1

Understanding Communication Failures

Pitre et al.

Befund 3'211 gefilterte fehlgeschlagene Mehr-Runden-Verläufe: Korrektur und Konsens dominieren in diesem Fehlerkorpus.

Ableitung Dissens, Belege und Minderheitensignale prüfen; keine Fehlerquote aller Läufe.

Quellen: Kim v3; Tieu v1; MAST v3 (NeurIPS 2025); Pitre v1. +80,8 / −70,0 % sind relative Änderungen, keine Prozentpunkte; kein organisationsweiter Produktivitätseffekt.

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

- **Aufgabe vor Anzahl:** Architektur je Aufgabe wählen; mehr Agenten können Leistung senken.
- **Prüfinstanzen:** Risikogerecht testen; keinen sicheren Zentralisierungseffekt annehmen.
- **Business Case:** Koordinationskosten und Verifikation einrechnen.
- **Kommunikation:** Regeln für Dissens und Korrektur festlegen.

StGallen imt Business School | arXiv

Quellen: arXiv

DEEP DIVE | EVALUATION UND GRENZEN

Längere Aufgabenhorizonte, begrenzte Sicherheitsevidenz

Historischer Aufgabenhorizont

METR (Kwa et al.) · TH1 · Claude 3.7 Sonnet

56 Min.

95-%-KI 28–94 Min. (TH1). Menschliche Expertendauer von Aufgaben, die das Modell zu 50 % löst; nicht Agentenlaufzeit.

≈ 207 Tage

Verdopplungszeit, historischer Fit 2019 bis Anfang 2025; 95-%-KI 166–240 Tage.

Szenario der Autoren (2025) Bei Trendfortsetzung um 2030 ausgewählte Softwareaufgaben von ca. einem Monat menschlicher Referenzdauer. Keine Planungszusage.

Was Sicherheitsprüfungen leisten

Einzelstudien aus dem Dantas-Review · Originalbefunde je Zeile

24,8 % bzw. 34,2 % (zwei Settings)

Planetarium: semantisch korrektes STRIPS-PDDL; GPT-4o zero-shot bzw. bestes feinabgestimmtes Gemma-2-27B-IT.

40,63 → 14,07 %

ProbGuard (Reflect; SafeAgentBench, GPT-4o-mini, θ = 0,9): −65,37 % unsichere Taskausgänge rel.; Completion 59,38 → 47,74 %.

bis zu 94,1 %

VerifierTax (τ-bench Retail, GPT-OSS-20B): abgefangene regelwidrige Toolvorschläge – Nenner Vorschläge, nicht Aufgaben.

< 5 %

sicher UND erfolgreich erledigte Aufgaben in den meisten Settings (GLM-4-9B, GPT-OSS-20B).

Review: 38 Studien; 7/9 Kernaussagen «sehr niedrig» im angepassten GRADE-Autorenrating.

Zusammenarbeit mit Fachleuten

Hayati et al. (2026) · arXiv:2610.21109

115+	50	40
Stunden Aufnahmen	Fachpersonen	Berufe

Befund Kolab testet, ob Modelle Übernehmen/Ändern/Ablehnen und subjektive Qualitätsurteile der Fachleute vorhersagen; sie überschätzen die Zusammenarbeit.

Grenze Mehr Kontext half diesen Vorhersagen nicht konsistent; keine kausale Produktivitätsmessung.

Quellen: METR TH1.1-Blog (29.01.2026, TH1); Kwa v4; Planetarium; ProbGuard; VerifierTax; Dantas v3; Kolab v1. Unterschiedliche Messgrössen und Nenner. VerifierTax-Protokoll: nach drei erfolglosen Revisionen letzter Vorschlag angenommen.

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

- **Autonomie:** An gemessene Zuverlässigkeit und Prüfkapazität koppeln.
- **Zielgrösse:** Zusammenarbeit mit Fachleuten, nicht nur Automatisierung.
- **Sicherheit:** An Evidenzqualität messen, nicht nur an Blockierraten.
- **Szenarien:** Extrapolationen (Monate, Verdopplung) nicht als Planbasis nutzen.

StGallen imt Business School | arXiv

Quellen: arXiv

DEEP DIVE | AUSBLICK

Fünf Prüfhypothesen für die weitere Agentenentwicklung

Synthese des KI-Instituts (StGallen imt Business School): Prüfhypothesen, keine Prognosen der zitierten Autor:innen und keine gesicherte empirische Vorhersage

01

Längere Aufgabenhorizonte

Historischer METR-Fit 2019–Anfang 2025: Verdopplung ca. alle 207 Tage (≈ 7 Monate). Kein fortlaufendes Gesetz; Zukunft nur als bedingtes Szenario.

Kwa et al. (v4); METR TH1.1, 2026

02

Selbstverbessernde Loops

Agentencode und Harness iterativ anpassen; Validierung und Audits testen. Kapazitätskontrolle: PAC-Theorie.

Zhang et al., 2025; Ren et al., 2026; C. L. Wang et al., 2025; X. Wang et al., 2026

03

Graphen nach Mass

Aufgabenspezifische Workflow-Graphen testen; kein belegter allgemeiner Ersatz fixer Agenten-Teams.

Kim et al., 2025; Tieu et al., 2026

04

Robuste Kommunikation

Dissens und Korrektur prüfen: Pitre analysiert gefilterte Fehlverläufe, nicht alle Agentenfehler.

Cemri et al., 2025; Pitre et al., 2026

05

Prüfbare Sicherheit

Forschungsagenda: Sicherheit ganzer Pläne und Vorhersage menschlicher Nutzung prüfen.

Dantas et al., 2026; Hayati et al., 2026

Quellen: wie angegeben; Fassungen und Publikationsstatus auf Folien 12 und 21–23; Stand 7. November 2026.

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

- **Hypothesen:** Die fünf Linien prüfen, nicht als Fahrplan behandeln.
- **Messung:** Fortschritt an eigenen Aufgaben und Prüfprotokollen messen.
- **Piloten:** Jede Linie mit Pilot und klarem Abbruchkriterium testen.
- **Trennung:** Autorenaussagen und Institutseinschätzungen getrennt halten.

StGallen imt Business School | arXiv

Quellen: arXiv

DEEP DIVE | IMPLIKATIONEN

Implikationen für das Management: sechs Ableitungen

Ableitungen des KI-Instituts (StGallen imt Business School): Empfehlungen zum Testen, keine Wirkungsbelege oder universellen Freigabenormen

Architektur nach Aufgabe wählen

PlanCraft: unabhängige Agenten bis −70,0 % relativ zur Einzelagent-Baseline. Kein allgemeiner Planungseffekt.

Kim et al., 2025

Prüfinstanzen risikogerecht testen

Fehlerfortpflanzung hängt von der Architektur ab; kein isolierter kausaler Nachweis zentraler Verifikation.

Kim et al., 2025; Cemri et al., 2025

Verantwortung auf Systemebene

Konzeptionelles CI/CD-Modell: Übersteigt Output die Prüfkapazität, können Freigaben ritualisieren. Annahmen, keine Rechtsaussage.

Romanchuk & Bondar, 2026

Selbstverbesserung mit Leitplanken

Empfehlung: unabhängig validieren und Harness auditieren. Statistische Kapazitätskontrolle gilt im PAC-Modell.

C. L. Wang et al., 2025; X. Wang et al., 2026

Zusammenarbeit messen

Eigene Aufgaben testen: Übernehmen, Ändern, Verwerfen durch Fachleute neben Qualität und Aufwand messen.

Hayati et al., 2026 (Zusammenarbeit); Kwa et al., 2025 (Aufgabenhorizont)

Sicherheitsversprechen prüfen

Dantas: 7/9 Kernaussagen «sehr niedrig» im angepassten GRADE-Autorenrating; keine universelle Wirkungsdauer.

Dantas et al., 2026

SO WHAT? | SICHT DES KI-INSTITUTS, STGALLEN IMT BUSINESS SCHOOL

Einordnung und Empfehlungen des Instituts für Artificial Intelligence in Business and Management · keine Aussagen der zitierten Autor:innen

- **90 Tage:** Mandat, Architekturwahl und Prüfinstanz schriftlich festlegen.
- **Eigene Tests:** Zusammenarbeit und Sicherheit selbst messen.
- **Leitplanken:** Selbstverbesserung und Multi-Agent nur kontrolliert freigeben.
- **Verantwortung:** So zuschneiden, dass Freigabe kein Ritual wird.

StGallen imt Business School

AUSBLICK | EMPFEHLUNGEN FÜR 12–24 MONATE

Fünf Entscheidungen für Führungsgremien

01 Mandate und Autonomiestufen festlegen

Autonomie nach Tragweite, Risiko, Umkehrbarkeit sowie Zielen und Werten begrenzen. Rolle, Rechte und Freigaben festlegen. Die Letztverantwortung verbleibt bei den zuständigen Menschen und Organen.

02 Managementarchitektur und Harness gemeinsam gestalten

In Menschen, Organisation, Kontext, Workflows und Evaluation investieren; Abhängigkeit von einem Modellanbieter begrenzen.

03 Wirkung auf Unternehmensebene messen

Tinkoff et al. (2026): 80 % berichten höhere eigene Produktivität, 37 % einen positiven EBIT-Beitrag durch KI. Anteile der Befragten, keine Effektgrössen. Eigene Wirkung messen.

04 Fähigkeiten und Nachwuchs sichern

Fachwissen, Urteilskraft und Nachwuchs gezielt entwickeln. Aufgabenprofile und Lernwege im eigenen Unternehmen beobachten; aus Beschäftigungstrends keinen kausalen KI-Effekt ableiten.

05 Agenten-Loops mit Prüfpunkten steuern

Risikogerechte Prüfpunkte testen; ein allgemeiner kausaler Vorteil zentraler Verifikation ist nicht belegt (Kim et al.).

Quellen: Tinkoff et al. (2026), McKinsey & Company, 25.08.2026, Abb. 5 und Abschnitt zu EBIT; Gesamtumfrage n = 1'719, 97 Länder, 04.05.–08.06.2026, BIP-gewichtet; Selbstauskunft, kein kausaler Wirkungsnachweis. Abegglen et al. (2026), Abschn. IV.5; Kim et al. v3. Empfehlungen dieser Präsentation.



KERNAUSSAGEN

Menschen.
Organisation.
Verantwortete KI.

Management vor Technologie.

StGallen imt Business School

Die Zukunft ist eine Managemententscheidung

Wert entsteht im hybriden Managementsystem.

Menschen, Organisation und Harness wirken zusammen. Das St. Galler Konzept verbindet normative, strategische und operative Führung mit Strukturen, Aktivitäten und Verhalten.

NORMATIV

Zweck und Werte

Agenten bereiten Urteile vor; nicht delegierbar: Werteabwägung, Folgenverantwortung.

STRATEGISCH

Fähigkeiten

Erfolgspotenziale, Fähigkeiten, Architekturwahl und Abhängigkeiten.

OPERATIV

Mandate im Alltag

Workflows, Rollen, Rechte, Prüfungen und Lernen.

Kohärenz: Passen Strukturen, vorgesehene Aufgaben und tatsächliches Verhalten zusammen?

Diese Woche ein Mandat klären: Zweck, Autonomie, Belege und Eskalation.

Eigene Managementsynthese nach Abegglen & Bleicher (2021) und Abegglen et al. (2026, TUM-I26129), Abschn. II–V. Agentonomics: Forschungs- und Gestaltungsprogramm; Nutzen, Grenzen und Wirtschaftlichkeit sind empirisch zu prüfen.



Die Lücke zwischen breiter Nutzung und schmaler Wirkung schliesst sich dort, wo Unternehmen Workflows neu gestalten, Agenten klar regeln und Ergebnisse überprüfen.»

DAVID WANG, SYNTHESE FÜR DIESE PRÄSENTATION, NOVEMBER 2026

StGallen imt Business School

STGALLEN IMT BUSINESS SCHOOL RESEARCH

Autoren und ausgewählte Quellen



Prof. Dr. Christian Abegglen
Verwaltungsratspräsident der StGallen imt Business School



David Wang
Direktor, Institut für Artificial Intelligence in Business and Management
Senior Lecturer, StGallen imt Business School

Institut für Artificial Intelligence in Business and Management

Schwerpunkte: Prozesstransformation, Mensch-KI-Zusammenarbeit, KI-gestützte Entscheidungen, AI Governance.

Die StGallen imt Business School wurde 2020 als Management- und Technologie-Spin-off initiiert und ist seit 2022 eine eigenständige private Business School mit Sitz in St. Gallen. Ihre Programme richten sich an Executives, C-Level und Unternehmer.

AUSGEWÄHLTE QUELLEN | VOLLSTÄNDIGES VERZEICHNIS AUF FOLIE 21 BIS 23

- Abegglen, C., & Bleicher, K. (2021). Das Konzept Integriertes Management (10. Aufl.). Campus.
- Abegglen, C. (2026). Agentonomics auf der Grundlage des St. Galler Konzepts: 38 Thesen (Arbeitspapier 001-AGE). St. Galler Gesellschaft für Integriertes Management.
- Abegglen, C., Knoll, A. C., Suckfüll, H. C., & von Grone, J. (2026). Agentonomics: Management im Zeitalter intelligenter Agentensysteme (TUM-I26129).
- Artificial Analysis. (2026, 27. Oktober). Announcing Artificial Analysis Intelligence Index v5; (o. D.). Changelog.
- Kim, Y. et al. (2025). Towards a science of scaling agent systems (arXiv:2512.08296). arXiv.
- Model Context Protocol. (2026). Specification (Version 2026-07-28).
- Tinkoff, D., Van der Veken, L., Chui, M., & Balakrishnan, T. (2026). The state of AI in 2026: On the road to ROI. McKinsey & Company.
- Wang, D. (2026). Integriertes Management in Zeiten von AI [verwandter Forschungsartikel]. StGallen imt Business School.

Verwendete Quellen (1/3)

40 verwendete + 22 ergänzende und weiterführende Quellen = 62 Einträge. Letzte Quellenprüfung: 8. November 2026, Ergänzung 10. November 2026; Benchmark-Aufnahme: 7. November 2026.
Ergänzende Quellen (22): Artikel «Von Chatbots zu Agenten, Loops und Graphen» (StGallen imt Business School, November 2026), Abschnitt «Quellen – Ergänzende und weiterführende Quellen».

Abegglen, C., & Bleicher, K. (2021). Das Konzept Integriertes Management: Visionen – Missionen – Programme (10., aktualisierte und erweiterte Aufl.). Campus. Abgerufen am 8. November 2026, von https://campus.de/business-management/fuehrung-management/das-konzept-integriertes-management/CAM51171

Abegglen, C. (2026). Agentonomics auf der Grundlage des St. Galler Konzepts: Konzeptionelle Vorarbeiten. 38 Thesen zur Gestaltung hybrider Managementsysteme auf der Grundlage des St. Galler Konzepts Integriertes Management (Arbeitspapier Nr. 001-AGE; veröffentlicht 31. Juli 2026, Stand 10. September 2026). St. Galler Gesellschaft für Integriertes Management. Abgerufen am 7. November 2026, von https://gimsg.ch/publikationen/agentonomics-konzeptionelle-vorarbeiten

Abegglen, C., Knoll, A. C., Suckfüll, H. C., & von Grone, J. (2026). Agentonomics: Management im Zeitalter intelligenter Agentensysteme. Grundzüge eines managementwissenschaftlichen Forschungs- und Gestaltungsansatzes (Technischer Bericht TUM-I26129). Technische Universität München, Institut für Informatik. Abgerufen am 8. November 2026, von https://gimsg.ch/publikationen/agentonomics-management-intelligente-agentensysteme

Aider. (o. D.). LLM leaderboards [Polyglot-Benchmark, 225 Aufgaben]. Abgerufen am 7. November 2026, von https://aider.chat/docs/leaderboards/

Anthropic. (2024, 19. Dezember). Building effective agents. Abgerufen am 7. November 2026, von https://www.anthropic.com/engineering/building-effective-agents

Anthropic. (2025, 26. November). Effective harnesses for long-running agents. Abgerufen am 7. November 2026, von https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents

Anthropic. (2026, 9. November). Introducing Claude Fable 5.2 and Claude Mythos 5.2. Abgerufen am 10. November 2026, von https://www.anthropic.com/claude-fable-and-mythos-5-2

Artificial Analysis. (o. D.). Changelog. Abgerufen am 7. November 2026, von https://artificialanalysis.ai/changelog

Artificial Analysis. (o. D.). Intelligence Index v5.0.1 [Screenshot/Diagrammexport; Screenshot: 07.11.2026]. Abgerufen am 7. November 2026, von https://artificialanalysis.ai/models#intelligence

Artificial Analysis. (o. D.). Methodology: Intelligence benchmarking. Abgerufen am 7. November 2026, von https://artificialanalysis.ai/methodology/intelligence-benchmarking

Artificial Analysis. (2026, 27. Oktober). Announcing Artificial Analysis Intelligence Index v5. Abgerufen am 7. November 2026, von https://artificialanalysis.ai/articles/artificial-analysis-intelligence-index-v5

Artificial Analysis. (2026, 9. November). Changelog [Einträge zu Claude Fable 5.2; Intelligence Index 65 (max)]. Abgerufen am 10. November 2026, von https://artificialanalysis.ai/changelog

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2025). Why do multi-agent LLM systems fail? (arXiv:2503.13657v3) [gelesene Fassung v3; auch NeurIPS 2025, Datasets and Benchmarks Track, gleiche 13 Autor:innen]. arXiv. https://arxiv.org/abs/2503.13657v3

Chen, M., et al. (2021). Evaluating large language models trained on code (arXiv:2107.03374) [HumanEval-Definition, 164 Aufgaben]. arXiv. https://arxiv.org/abs/2107.03374

Verwendete Quellen (2/3)

40 verwendete + 22 ergänzende und weiterführende Quellen = 62 Einträge. Letzte Quellenprüfung: 8. November 2026, Ergänzung 10. November 2026; Benchmark-Aufnahme: 7. November 2026.
Ergänzende Quellen (22): Artikel «Von Chatbots zu Agenten, Loops und Graphen» (StGallen imt Business School, November 2026), Abschnitt «Quellen – Ergänzende und weiterführende Quellen».

Dantas, P., Cordeiro, L., Nowroozi, E., & Tihanyi, N. (2026). Toward safe LLM agents: A survey of specification, verification, and enforcement (arXiv:2608.14590v3) [Preprint, gelesene Fassung v3; laut arXiv-Kommentar in Begutachtung bei ACM Computing Surveys]. arXiv. https://arxiv.org/abs/2608.14590v3

Hayati, S. A., Setiani, P. A., Wang, R., Li, Z., Le, K. C., Lee, Y.-J., & Kang, D. (2026). Kolab: Benchmarking human-agent collaboration in long-horizon workflows (arXiv:2610.21109v1) [Preprint]. arXiv. https://arxiv.org/abs/2610.21109v1

Kim, Y., Gu, K., Park, C., Park, C., Schmidgall, S., Heydari, A. A., Yan, Y., Zhang, Z., Zhuang, Y., Liu, Y., Malhotra, M., Liang, P. P., Park, H. W., Yang, Y., Xu, X., Du, Y., Patel, S., Althoff, T., McDuff, D., & Liu, X. (2025). Towards a science of scaling agent systems (arXiv:2512.08296v3) [Preprint, gelesene Fassung v3]. arXiv. https://arxiv.org/abs/2512.08296v3

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Painter, C., … Chan, L. (2025). Measuring AI ability to complete long software tasks (arXiv:2503.14499v4) [Preprint, gelesene Fassung v4 (26 Autor:innen)]. arXiv. https://arxiv.org/abs/2503.14499v4

McKinsey & Company. (2025, 5. November). The state of AI in 2025: Agents, innovation, and transformation [Global Survey; n = 1'993, Feldzeit 25.06.–29.07.2025]. Abgerufen am 7. November 2026, von https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2025

METR. (2026, 29. Januar). Time Horizon 1.1. Abgerufen am 7. November 2026, von https://metr.org/blog/2026-01-29-time-horizon-1-1/

Model Context Protocol. (2026). Specification (Version 2026-07-28). Abgerufen am 7. November 2026, von https://modelcontextprotocol.io/specification/2026-07-28

Pitre, P., Hou, J., Fang, Y., Ramakrishnan, N., & Wang, X. (2026). Understanding communication failures in multi-agent LLM systems (arXiv:2610.32060v1) [Preprint]. arXiv. https://arxiv.org/abs/2610.32060v1

Ren, Z., Chen, Y., Guo, D., Rong, G., Li, T., Xiong, R. B., Lan, Q., Wang, W., Li, N., Yang, Y., Zhuge, M., & Schmidhuber, J. (2026). Self-improvements in modern agentic systems: A survey (arXiv:2607.13104v1) [Preprint]. arXiv. https://arxiv.org/abs/2607.13104v1

Romanchuk, O., & Bondar, R. (2026). The responsibility vacuum: Organizational failure in scaled agent systems (arXiv:2601.15059v1) [Preprint]. arXiv. https://arxiv.org/abs/2601.15059v1

Sah, T., Srivastava, V., Sah, D., & Jordan, K. (2026). The verifier tax: Horizon dependent safety–success tradeoffs in tool using LLM agents (arXiv:2603.19328v1) [Preprint]. arXiv. https://arxiv.org/abs/2603.19328v1

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning (arXiv:2303.11366v4) [gelesene arXiv-Fassung v4, sechs Autor:innen]. arXiv. https://arxiv.org/abs/2303.11366v4

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning [Konferenzfassung, fünf Autor:innen]. NeurIPS 2023. https://neurips.cc/virtual/2023/poster/70114

StGallen imt Business School

ANHANG | LITERATUR IN ANLEHNUNG AN APA 7

Verwendete Quellen (3/3)

40 verwendete + 22 ergänzende und weiterführende Quellen = 62 Einträge. Letzte Quellenprüfung: 8. November 2026, Ergänzung 10. November 2026; Benchmark-Aufnahme: 7. November 2026.
Ergänzende Quellen (22): Artikel «Von Chatbots zu Agenten, Loops und Graphen» (StGallen imt Business School, November 2026), Abschnitt «Quellen – Ergänzende und weiterführende Quellen».

StGallen imt Business School. (o. D.-a). Über StGallen imt. Abgerufen am 7. November 2026, von https://stgallenimt.ch/support_aboutimt.html

StGallen imt Business School. (o. D.-c). Institut für Artificial Intelligence in Business and Management. Abgerufen am 7. November 2026, von https://stgallenimt.ch/institut_ai_business_management.html

StGallen imt Business School. (o. D.-e). Netzwerk, Partner & Fakultät. Abgerufen am 7. November 2026, von https://stgallenimt.ch/netzwerk.html

St. Galler Gesellschaft für Integriertes Management. (o. D.). Projekt Agentonomics. Abgerufen am 7. November 2026, von https://gimsg.ch/projekt-agentonomics

Tieu, K., Fu, D., Xia, Y., Li, H., Yan, H., & He, J. (2026). Inference-time graph engineering for multi-agent LLM workflows (arXiv:2609.05774v1) [Preprint]. arXiv. https://arxiv.org/abs/2609.05774v1

Tinkoff, D., Van der Veken, L., Chui, M., & Balakrishnan, T. (2026, 25. August). The state of AI in 2026: On the road to ROI [Global Survey; n = 1'719]. McKinsey & Company. Abgerufen am 7. November 2026, von https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Wang, C. L., Dorchen, K., & Jin, P. (2025). On the statistical limits of self-improving agents (arXiv:2510.04399v3) [gelesene Preprint-Fassung v3; veröffentlicht in Transactions on Machine Learning Research, 2026]. arXiv. https://arxiv.org/abs/2510.04399v3

Wang, D. (2026). Integriertes Management in Zeiten von AI [verwandter Forschungsartikel]. StGallen imt Business School. Abgerufen am 7. November 2026, von https://stgallenimt.ch/research/integriertes-management-in-zeiten-von-ai.html

Wang, H., Poskitt, C. M., Wei, J., & Sun, J. (2025). ProbGuard: Proactive runtime monitoring for LLM agent safety via probabilistic prediction (arXiv:2508.00500v4) [gelesene Fassung v4; laut arXiv-Fassung ASE '26]. arXiv. https://arxiv.org/abs/2508.00500v4

Wang, X., Zhang, X., & Shao, J. (2026). Auditing harness tampering in self-improving agents (arXiv:2609.00069v1) [Preprint]. arXiv. https://arxiv.org/abs/2609.00069v1

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models [ICLR 2023; gelesene arXiv-Fassung v3 = Camera-ready laut arXiv-Kommentar]. https://arxiv.org/abs/2210.03629v3

Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Gödel machine: Open-ended evolution of self-improving agents (arXiv:2505.22954v3) [gelesene Fassung v3; Konferenzfassung ICLR 2026]. arXiv. https://arxiv.org/abs/2505.22954v3

Zuo, M., Piedrahita Velez, F., Li, X., Littman, M. L., & Bach, S. H. (2025). Planetarium: A rigorous benchmark for translating text to structured planning languages. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the ACL: Human Language Technologies (Bd. 1, S. 11223–11240). https://doi.org/10.18653/v1/2025.naacl-long.560