# Comparing Notes on Trustworthy AI 2025

## Part 1: AI Agents 101. What They Are, How They Work, and How They Can Create Value

***German version below***

**Host:** appliedAI Institute for Europe
**Date:** 6 May 2025
**Location:** House of Communication, Munich

The first event of the 2025 series brought together four practitioners to cut through the hype around AI agents: examining what they actually are, where they create real value today, what organisations need to get started, and how to implement them responsibly. The discussion was grounded in concrete use cases from pharma, security, and legal operations, and aimed at Bavarian SMEs, startups, and public sector organisations taking their first steps.

---

## The panel

**Dr. Clara Neppel** is Senior Director at IEEE, bringing a global standards and governance perspective to the conversation. She set the stakes early: only 6% of German industry can currently measure the value added through data use, which frames both the opportunity and the gap that responsible AI deployment must bridge.

**Daniela Rittmeier** leads the Gen AI Accelerator at Capgemini, working on translating AI potential into business reality. She brought hands-on evidence from pharmaceutical regulatory compliance, warned against "POC purgatory," and pushed for integration over experimentation.

**Marieke Luise Merkle** is an Associated Partner at Noerr specialising in digital business law. She provided the legal lens throughout: on liability allocation between providers and deployers, contractual human-oversight requirements, and why early legal consultation is not optional.

**Maximilian Werk** is Head of Engineering at Jina AI, building the technical infrastructure that agentic systems run on. He grounded the conversation in architecture realities and offered the session's most direct advice: "Find your problem first. Easy: agents can open and read documents - start here."

The panel was moderated by **Claudia Baumgartner**, Trustworthy AI Expert at the appliedAI Institute for Europe gGmbH.

*Note: This document includes AI-generated summaries and translations created with the assistance of Gemini. / Dieses Dokument enthält KI-generierte Zusammenfassungen und Übersetzungen, die mit Unterstützung von Gemini erstellt wurden.*

# Discussion

## What are AI agents, and what are they not?

AI agents are autonomous systems that can perceive their environment, make decisions, and take actions to achieve specific goals. Unlike traditional AI systems limited to their training data, agents can interact with external systems, access online information, and perform tasks with varying levels of autonomy, including across multiple modalities such as text, images, and audio.

The panel was clear that agents are not chatbots: chatbots respond to prompts, while agents plan, decide, and act. Three misconceptions dominated the opening exchange: the overestimation of existential risk, the conflation of simple reactive systems with genuinely autonomous planners, and the assumption that entirely new governance structures must be invented before anything can be deployed.

## Where agents create real business value

Daniela Rittmeier described specialised agents in the pharmaceutical sector already handling regulatory compliance and market entry planning. Other live applications the panel covered included sustainability systems reducing food waste through smarter prediction, security operations coordinating emergency responses during blackouts, and legal departments where document-processing automation frees professionals for higher-level work.

AI agents deliver value across four dimensions: enhanced productivity through repetitive-task automation; cost reduction through streamlined processes; quality improvements from more consistent outputs; and better scalability, handling greater workloads without proportionally more resources.

The note of caution was equally concrete. Agents can produce unintended side effects, including mistakenly deleting data. From a security perspective, agents that can serve as defenders can, in adversarial conditions, also become attackers. Organisations must realistically assess both capabilities and limitations before deployment.

## Implementation requirements: getting started

Most organisations need four technical elements to begin: access to large language models (preferably on-premise rather than cloud-based), evaluation systems to measure performance, flexible architecture allowing agents to adapt, and standard building blocks reusable across projects.

"Think big, start small, show value" emerged as the recommended philosophy. Daniela was specific: integrate agents into existing infrastructure rather than creating isolated experiments. Iterative development with continuous learning cycles can show tangible results within 2 to 4 weeks. The agreed starting point for any organisation: identify a document-processing task such as email categorisation, form extraction, or meeting-note summaries, and automate it.

*Note: This document includes AI-generated summaries and translations created with the assistance of Gemini. / Dieses Dokument enthält KI-generierte Zusammenfassungen und Übersetzungen, die mit Unterstützung von Gemini erstellt wurden.*

## Risk, governance, and the legal dimension

Effective risk management requires human oversight with real intervention capabilities, transparency measures that make agent operations understandable, and systematic risk assessment for failure modes specific to agentic systems. Data privacy compliance, particularly with GDPR, and decision traceability for audits remain foundational requirements.

Practical guardrails recommended by the panel: designing systems with human-in-the-loop escalation layers, compartmentalising to limit agent access and capabilities, restricting agents to a bounded amount of change per day, and requiring multiple approvals for high-risk actions.

Marieke's legal framing was direct: early consultation is not optional, it is essential. 100% compliance is rarely achievable in initial deployments. The goal is realistic expectations, human-oversight requirements written explicitly into contracts, and documented oversight procedures that hold up under liability scrutiny. Liability must be clearly allocated between the provider and the deployer of the AI agent.

## Common pitfalls to avoid

"POC purgatory" topped the list: too many proofs-of-concept that stay in the pipeline and never generate real value. Centralised IT bottlenecks were the second failure mode - domain experts should lead implementation together with IT teams. Starting too large, with complex, high-risk, or customer-facing applications, was the third. Neglecting guardrails - clear boundaries, maximum code modifications per day, escalation triggers - rounded out the recurring mistakes.

On organisational readiness, the audience asked what capabilities are genuinely required. The consensus: decentralised implementation (hub-and-spoke rather than centralised IT), honest assessment of internal competencies, domain expertise combined with technical knowledge, and a strategy that protects valuable industrial data assets.

## Key takeaways

Implementing trustworthy AI agents means focusing on concrete value creation, responsible deployment, and appropriate risk management. The most successful approach combines:

- Clear problem definition before selecting or building an agent.
- Realistic assessment of capabilities and limitations upfront.
- Iterative development with continuous learning in cross-disciplinary teams.
- Appropriate human oversight built into the system, not bolted on later.
- Cross-functional collaboration between technical, business, and legal perspectives.
- Explicit liability definition in contracts between provider and deployer.

# Comparing Notes on Trustworthy AI 2025

## Teil 1: KI-Agenten 101. Was sie sind, wie sie funktionieren und wie sie einen Mehrwert schaffen können

**Gastgeber:** appliedAI Institute for Europe
**Datum:** 6. Mai 2025
**Ort:** Haus der Kommunikation, München

Die erste Veranstaltung der Reihe 2025 brachte vier Praktiker zusammen, um den Hype um KI-Agenten zu durchbrechen: Es wurde untersucht, was sie tatsächlich sind, wo sie heute echten Mehrwert schaffen, was Organisationen für den Einstieg benötigen und wie man sie verantwortungsvoll implementiert. Die Diskussion basierte auf konkreten Anwendungsfällen aus den Bereichen Pharma, Sicherheit und Rechtsabteilung und richtete sich an bayerische KMUs, Start-ups und Organisationen des öffentlichen Sektors, die ihre ersten Schritte unternehmen.

## Das Panel

**Dr. Clara Neppel** ist Senior Director bei IEEE und bringt eine globale Perspektive auf Standards und Governance in das Gespräch ein. Sie betonte früh die Herausforderungen: Nur 6 % der deutschen Industrie können derzeit den Mehrwert messen, der durch die Datennutzung entsteht, was sowohl die Chance als auch die Lücke verdeutlicht, die durch den verantwortungsvollen Einsatz von KI geschlossen werden muss.

**Daniela Rittmeier** leitet den Gen AI Accelerator bei Capgemini und arbeitet daran, das KI-Potenzial in die Geschäftsrealität umzusetzen. Sie lieferte praktische Beweise aus der Einhaltung pharmazeutischer Vorschriften, warnte vor dem "POC Purgatory" (Machbarkeitsstudien-Fegefeuer) und drängte auf Integration statt auf Experimente.

**Marieke Luise Merkle** ist Associated Partner bei Noerr und spezialisiert auf digitales Wirtschaftsrecht. Sie lieferte die rechtliche Perspektive durchgehend: zur Haftungsverteilung zwischen Anbietern und Betreibern, zu vertraglichen Anforderungen an die menschliche Aufsicht und warum eine frühzeitige Rechtsberatung nicht optional ist.

**Maximilian Werk** ist Head of Engineering bei Jina AI und baut die technische Infrastruktur, auf der agentenbasierte Systeme laufen. Er verankerte das Gespräch in architektonischen Realitäten und bot den direktesten Rat der Sitzung: "Finden Sie zuerst Ihr Problem. Ganz einfach: Agenten können Dokumente öffnen und lesen – fangen Sie hier an."

Das Panel wurde moderiert von **Claudia Baumgartner**, Trustworthy AI Expert bei der appliedAI Institute for Europe gGmbH.

# Diskussion

## Was sind KI-Agenten und was sind sie nicht?

KI-Agenten sind autonome Systeme, die ihre Umgebung wahrnehmen, Entscheidungen treffen und Handlungen ausführen können, um bestimmte Ziele zu erreichen. Im Gegensatz zu traditionellen KI-Systemen, die auf ihre Trainingsdaten beschränkt sind, können Agenten mit externen Systemen interagieren, auf Online-Informationen zugreifen und Aufgaben mit unterschiedlichem Grad an Autonomie ausführen, auch über mehrere Modalitäten wie Text, Bilder und Audio hinweg.

Das Panel stellte klar, dass Agenten keine Chatbots sind: Chatbots reagieren auf Prompts, während Agenten planen, entscheiden und handeln. Drei Missverständnisse dominierten den ersten Austausch: die Überschätzung des existenziellen Risikos, die Verwechslung einfacher reaktiver Systeme mit wirklich autonomen Planern und die Annahme, dass völlig neue Governance-Strukturen erfunden werden müssen, bevor überhaupt etwas eingesetzt werden kann.

## Wo Agenten echten Geschäftswert schaffen

Daniela Rittmeier beschrieb spezialisierte Agenten im Pharmasektor, die bereits die Einhaltung gesetzlicher Vorschriften und die Planung der Markteinführung übernehmen. Andere Live-Anwendungen, die das Panel behandelte, umfassten Nachhaltigkeitssysteme, die Lebensmittelverschwendung durch intelligentere Vorhersagen reduzieren, Sicherheitsoperationen, die Notfallreaktionen während Stromausfällen koordinieren, und Rechtsabteilungen, in denen die Automatisierung der Dokumentenverarbeitung Fachleute für höherwertige Arbeit freisetzt.

KI-Agenten liefern Mehrwert in vier Dimensionen: gesteigerte Produktivität durch Automatisierung sich wiederholender Aufgaben; Kostensenkung durch optimierte Prozesse; Qualitätsverbesserungen durch konsistentere Ergebnisse; und bessere Skalierbarkeit, indem größere Arbeitslasten ohne proportional mehr Ressourcen bewältigt werden.

Die Warnung war ebenso konkret. Agenten können unbeabsichtigte Nebenwirkungen verursachen, einschließlich der versehentlichen Löschung von Daten. Aus Sicherheitsperspektive können Agenten, die als Verteidiger dienen, unter gegnerischen Bedingungen auch zu Angreifern werden. Organisationen müssen sowohl Fähigkeiten als auch Einschränkungen vor dem Einsatz realistisch einschätzen.

## Implementierungsanforderungen: Der Einstieg

Die meisten Organisationen benötigen vier technische Elemente für den Anfang: Zugang zu großen Sprachmodellen (vorzugsweise On-Premise statt Cloud-basiert), Evaluierungssysteme zur Leistungsmessung, flexible Architektur, die Agenten zur Anpassung ermöglicht, und standardisierte, über Projekte hinweg wiederverwendbare Bausteine.

"Groß denken, klein anfangen, Mehrwert zeigen" etablierte sich als die empfohlene Philosophie. Daniela war spezifisch: Agenten in die bestehende Infrastruktur integrieren, anstatt isolierte Experimente zu schaffen. Iterative Entwicklung mit kontinuierlichen Lernzyklen kann innerhalb von 2 bis 4 Wochen greifbare Ergebnisse zeigen. Der vereinbarte Ausgangspunkt für jede Organisation: Identifizieren Sie eine Dokumentenverarbeitungsaufgabe wie E-Mail-Kategorisierung, Formular-Extraktion oder Zusammenfassungen von Besprechungsnotizen und automatisieren Sie diese.

## Risiko, Governance und die rechtliche Dimension

Effektives Risikomanagement erfordert menschliche Aufsicht mit echten Eingriffsmöglichkeiten, Transparenzmaßnahmen, die Agentenoperationen verständlich machen, und eine systematische Risikobewertung für spezifische Fehlermodi agentenbasierter Systeme. Die Einhaltung des Datenschutzes, insbesondere der DSGVO, und die Nachvollziehbarkeit von Entscheidungen für Audits bleiben grundlegende Anforderungen.

Praktische Schutzmaßnahmen, die vom Panel empfohlen wurden: Entwicklung von Systemen mit Human-in-the-Loop-Eskalationsstufen, Kompartimentierung zur Begrenzung des Zugriffs und der Fähigkeiten von Agenten, Beschränkung der Agenten auf eine begrenzte Änderungsmenge pro Tag und die Notwendigkeit mehrerer Genehmigungen für risikoreiche Aktionen.

Mariekes rechtliche Einordnung war direkt: Frühzeitige Beratung ist nicht optional, sondern unerlässlich. Eine 100%ige Konformität ist bei ersten Einsätzen selten erreichbar. Das Ziel sind realistische Erwartungen, explizit in Verträgen festgelegte Anforderungen an die menschliche Aufsicht und dokumentierte Aufsichtsverfahren, die einer Haftungsprüfung standhalten. Die Haftung muss klar zwischen dem Anbieter und dem Betreiber des KI-Agenten aufgeteilt werden.

## Häufige Fehler, die es zu vermeiden gilt

Das "POC Purgatory" (Machbarkeitsstudien-Fegefeuer) stand ganz oben auf der Liste: zu viele Machbarkeitsstudien, die in der Pipeline bleiben und nie echten Mehrwert generieren. Zentralisierte IT-Engpässe waren der zweite Fehlermodus – Fachexperten sollten die Implementierung zusammen mit IT-Teams leiten. Zu groß zu starten, mit komplexen, risikoreichen oder kundenorientierten Anwendungen, war der dritte. Die Vernachlässigung von Schutzmaßnahmen – klare Grenzen, maximale Code-Änderungen pro Tag, Eskalationstrigger – rundete die wiederkehrenden Fehler ab.

In Bezug auf die organisatorische Bereitschaft fragte das Publikum, welche Fähigkeiten wirklich erforderlich sind. Der Konsens: dezentralisierte Implementierung (Hub-and-Spoke statt zentralisierter IT), ehrliche Bewertung interner Kompetenzen, Domänenexpertise kombiniert mit technischem Wissen und eine Strategie, die wertvolle industrielle Datenbestände schützt.

## Wichtigste Erkenntnisse

Die Implementierung vertrauenswürdiger KI-Agenten bedeutet, sich auf konkrete Wertschöpfung, verantwortungsvollen Einsatz und angemessenes Risikomanagement zu konzentrieren. Der erfolgreichste Ansatz kombiniert:

- Klare Problemdefinition vor der Auswahl oder dem Bau eines Agenten.
- Realistische Einschätzung von Fähigkeiten und Einschränkungen im Vorfeld.
- Iterative Entwicklung mit kontinuierlichem Lernen in interdisziplinären Teams.
- Angemessene menschliche Aufsicht, die in das System integriert und nicht nachträglich angefügt wird.
- Funktionsübergreifende Zusammenarbeit zwischen technischer, geschäftlicher und rechtlicher Perspektive.
- Explizite Haftungsdefinition in Verträgen zwischen Anbieter und Betreiber.