Comparing Notes on Trustworthy AI 2025

Part 4: The Ethics of AI Agents: Navigating Between Innovation and Responsibility

German version below

Host: LMU Munich / MCML
Date: 25 September 2025
Location: Carl Friedrich von Siemens Stiftung, Munich

The fourth event deepened the ethical dimensions of AI agents. The central insight: responsible implementation requires balance - between technological progress and human values, between efficiency gains and protection of fundamental rights, and between innovation and societal responsibility. Expert impulses and interactive audience surveys shaped a wide-ranging discussion.

The panel

Prof. Dr. Lena Kastner is a philosopher specialising in the ethics of AI. She framed the discussion around unavoidable trade-offs: efficiency and comfort must not come at the cost of human dignity.

Prof. Dr. Gitta Kutyniok works on the mathematical foundations and trustworthiness of AI. She highlighted fundamental problems—hallucinations, inaccuracies, enormous power consumption—and the necessity of mathematical foundations for AI systems.

Prof. Dr. Barbara Plank is a computational linguist specialising in language variation. Her critical finding: only 1% of world languages are supported by AI. Trustworthy AI agents must accept ambiguity.

The panel was moderated by **Johannes Büchs** .

Discussion

The Sorcerer's Apprentice moment

The discussion focused on central ethical tensions, such as regulation vs. free market, and AI autonomy vs. human agency. A practical recommendation for users was to consider not only what they get from AI, but also what the AI takes from them.

Trustworthiness and its complexity

The panel found no unified definition of trustworthy AI. The key principle that emerged was calibrated trust—rather than blind trust. There is a need for AI certification

systems, analogous to organic labels, as established quality standards are currently missing.

Societal impact and risk

Demographic pressures are forcing organisations to use AI. Systemic risks were illustrated by scenarios like the mass flight compensation scenario. Concerns were raised about selective news manipulation, the reinforcement of filter bubbles, and the potential undermining of democratic processes.

Critical ethical boundaries

There was clear expert consensus on unacceptable applications, such as "killer robots". The use of AI therapists was identified as problematic because therapy's goal is independence, not permanent AI dependence. The central trade-off of improving efficiency versus the loss of human expertise was discussed.

Europe vs. USA and democratic navigation

The European focus on safety may hinder growth compared to the faster progress and economic gain in the USA. The panel highlighted the opportunity for society to still make conscious decisions to democratically shape the AI future.

Key takeaways

The most successful approach requires:

- Awareness of unavoidable trade-offs: no technological solution without costs.
- Calibrated trust: gradual building through experience, instead of blind trust or blanket rejection.
- Inclusive and fair development: considering all 7,000 world languages as a long-term goal.
- Clear ethical boundaries: consensus on unacceptable applications and protection of human agency.
- Multidisciplinary approach: no single discipline has all the answers.

Core message: there is no magic without a price. The time for these decisions is now.

Comparing Notes on Trustworthy AI 2025

Teil 4: Die Ethik von KI-Agenten: Navigation zwischen Innovation und Verantwortung

English version above

Gastgeber: LMU München / MCML
Datum: 25. September 2025
Ort: Carl Friedrich von Siemens Stiftung, München

Das vierte Event vertiefte die ethischen Dimensionen von KI-Agenten. Die zentrale Erkenntnis: verantwortungsvolle Implementierung erfordert Gleichgewicht – zwischen technologischem Fortschritt und menschlichen Werten, zwischen Effizienzgewinnen und dem Schutz fundamentaler Rechte.

Das Panel

Prof. Dr. Lena Kästner ist Philosophin mit Spezialisierung auf KI-Ethik. Sie rahmte die Diskussion um unvermeidliche Trade-offs: Effizienz und Komfort dürfen nicht auf Kosten der menschlichen Würde gehen.

Prof. Dr. Gitta Kutyniok arbeitet an den mathematischen Grundlagen und der Vertrauenswürdigkeit von KI. Sie hob fundamentale Probleme hervor – Halluzinationen, Ungenauigkeiten, enormer Stromverbrauch – und die Notwendigkeit mathematischer Fundamente.

Prof. Dr. Barbara Plank ist Computerlinguistin mit Schwerpunkt Sprachvariation. Ihr kritischer Befund: Nur 1% aller Weltsprachen werden von KI unterstützt.

Das Panel wurde moderiert von **Johannes Büchs** .

Diskussion

Der Zauberlehrling-Moment

Die Diskussion beleuchtete zentrale ethische Spannungsfelder, wie Regulierung vs. freier Markt und KI-Autonomie vs. menschliche Handlungsfähigkeit. Eine praktische Empfehlung des Panels war, nicht nur zu bedenken, was man von KI bekommt, sondern auch, was die KI von einem nimmt.

Vertrauenswürdigkeit und ihre Komplexität

Das Panel fand keine einheitliche Definition vertrauenswürdiger KI. Das Konzept des kalibrierten Vertrauens – statt blindem Vertrauen – ist das Schlüsselprinzip. KI-Zertifizierungssysteme, analog zu Bio-Siegeln, sind erforderlich.

Gesellschaftliche Auswirkungen und Risiken

Organisationen sind aufgrund demografischen Drucks gezwungen, KI-Agenten einzusetzen. Das Massenflugverfahren-Szenario zeigte systemisches Risiko auf. Die Demokratie ist durch selektive Nachrichtenmanipulation und Filterblasen gefährdet.

Kritische ethische Grenzen

Es gab einen klaren Konsens über inakzeptable Anwendungen wie „keine Killerroboter". Die Nutzung von KI-Therapeuten ist problematisch, da das Therapieziel Unabhängigkeit ist. Das Flugzeugreparatur-Beispiel verdeutlicht den Trade-off: Effizienzgewinn vs. Verlust von Fachwissen.

Europa vs. USA und demokratische Navigation

Der Sicherheitsfokus Europas könnte das Wachstum im Vergleich zum schnellen Fortschritt in den USA behindern. Die Gesellschaft hat noch die Chance, die KI-Zukunft demokratisch zu gestalten.

Wichtigste Erkenntnisse

- Kein magischer Trick ohne Kosten – jede technologische Lösung hat Kompromisse.
- Kalibriertes Vertrauen statt blindes Vertrauen oder pauschale Ablehnung.
- Inklusive Entwicklung: Berücksichtigung aller 7,000 Weltsprachen als langfristiges Ziel.
- Klare ethische Grenzen: Konsens über inakzeptable Anwendungen.
- Multidisziplinäre Herangehensweise: keine einzelne Disziplin hat alle Antworten.

Kernbotschaft: Es gibt keine Magie ohne Preis. Die Zeit für diese Entscheidungen ist jetzt.