

Lernen in komplexen Systemen

KI, Human Factors und die Zukunft der Gefahrenabwehr

Von Thilo Stadelmann

Abstract: Moderne KI-Systeme erzielen bemerkenswerte Leistungen durch grundlegend stochastische Prozesse, basierend auf Machine-Learning-Modellen, die als hochdimensionale Wahrscheinlichkeitsdichtefunktionen fungieren und die wahrscheinlichsten Vorhersagen auf der Grundlage von Trainingsdaten ausgeben. Solche Systeme erreichen oder übertreffen im Durchschnitt die menschliche Leistung, ihre Methodik führt jedoch zu grundlegend anderen Fehlermodi als das menschliche Denken. Dies bedingt, dass Fehler auftreten werden, die aus menschlicher Sicht unsinnig erscheinen, zwar aufgrund ihrer probabilistischen Natur vorhersehbar wären, aber für Menschen unerwartet kommen.

Dies hat kritische Folgen in Umgebungen wie sicherheitsrelevanten Anwendungen, in denen Entscheidungen nicht rückgängig gemacht werden können. Denn KI-Systeme übernehmen Vorurteile, können Plausibilität nicht von faktischer Richtigkeit unterscheiden und zeigen selbst dann selbstsicheres Verhalten, wenn sie falsch liegen. Entsprechend müssen für den effektiven Einsatz von KI in Szenarien mit weitreichenden Folgen Prozesse implementiert werden, die sicherstellen, dass alle menschlichen Interessengruppen sich dieser Tatsachen bewusst sind, eine angemessene Skepsis gegenüber dem KI-System entwickeln und aktiv in die Entscheidungsfindung eingebunden bleiben. Speziell für militärische Anwendungen zeigt diese Erkenntnis, dass eine effektive Zusammenarbeit zwischen Mensch und KI mehr als nur Aufsicht erfordert: Sie erfordert Co-Learning-Frameworks, die eine sinnvolle menschliche Kontrolle durch bidirektionalen Informationsfluss gewährleisten. Wir geben einen Ausblick auf dezentrale, co-gelernte KI-Systeme, die durch gemeinsames Training in speziellen Co-Learning-Umgebungen auf einzelne Teams zugeschnitten wurden, um Machtkonzentrationsrisiken zu mindern und gleichzeitig wesentliche menschliche Fähigkeiten zu erhalten, darunter das moralische Urteilsvermögen, Gnade zu gewähren.

Biographie: Thilo Stadelmann ist Professor für Künstliche Intelligenz und Maschinelles Lernen an der ZHAW School of Engineering in Winterthur/Schweiz sowie Gründer und Leiter des dortigen Centre for Artificial Intelligence. Als Informatiker promovierte er an der Philipps-Universität Marburg und war vor seiner Berufung an die ZHAW mehrere Jahre in Fach- und Führungsfunktionen in der Automobilindustrie tätig. Er ist (Mit-)Gründer und Mitglied der Führungsspitze mehrerer Organisationen im Digitalbereich. Seine aktuellen Forschungsinteressen umfassen robustes Representation Learning sowie die gesellschaftlichen Auswirkungen von KI; seine Arbeiten wurden vielfach ausgezeichnet. Auf TEDx erklärt er "How not to fear AI"; sein Buch "Applied Data Science - Lessons Learned for the Data-Driven Business" erschien im Springer-Verlag.

Literatur: Thilo Stadelmann, Philipp H. Merkt, and Kasey Barr. The stochastic nature of machine learning and its implications for high-consequence AI. In: *AI and Ethics*, 6, 195, Springer, March 15, 2026. DOI <https://doi.org/10.1007/s43681-026-01042-1>.