SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Wadern, 9. Juli 2025

Schloss Dagstuhl:

Lässt sich KI beaufsichtigen?

Interdisziplinäres Seminar zu wirksamer menschlicher Aufsicht von KI-Systemen in Schloss Dagstuhl

Künstliche Intelligenz (KI) trifft zunehmend Entscheidungen, die tief in unser Leben eingreifen – von der medizinischen Diagnostik bis zur autonomen Mobilität. Zu hoffen, dass dabei immer alles gut geht, ist naiv. Menschen können zu Schaden kommen oder ihre Rechte verletzt werden, wenn KI-Systeme falsche Diagnosen vorschlagen oder diskriminierende Entscheidungen treffen. Nicht immer lassen sich diese Risiken technisch vollständig kontrollieren. Wirksame menschliche Aufsicht ist daher ein zentraler Baustein für den vertrauenswürdigen Einsatz von KI – und wird für besonders risikoreiche Systeme von der Europäischen KI-Verordnung zwingend verlangt. Diese ist im August 2024 in Kraft getreten, und die meisten Bestimmungen darin werden mit dem 2. August 2026 scharf geschaltet. Jetzt ist daher die Zeit, konkret zu werden bei der Frage: Was sind die Bedingungen für wirksame menschliche Aufsicht über KI?

Mit diesen Fragen beschäftigte sich das Dagstuhl-Seminar **„Challenges of Human Oversight: Achieving Human Control of AI-Based Systems“**, das vom **29. Juni bis 4. Juli 2025** auf Schloss Dagstuhl stattfand. Eingeladen waren internationale Expertinnen und Experten aus Informatik, Psychologie, Recht, Ethik, Kognitionswissenschaften und Technikgestaltung.

Zentrale Themen des Workshops waren Fragen rund um Wirksamkeit, Gerechtigkeit, Effektivität und Effizienz menschlicher Aufsicht. Intensiv diskutiert wurde beispielsweise, wie man Situationen vermeidet, in denen es viel einfacher ist, sich einer durch die KI vorgeschlagenen Entscheidung anzuschließen, als den Aufwand für einen gut begründeten Widerspruch aufzubringen.

Von der Forderung zur Umsetzung

Begriffe wie Human-in-the-Loop oder menschliche Aufsicht sind mittlerweile feste Bestandteile politischer Diskurse. Dennoch fehlt bislang ein konsistentes, interdisziplinäres Verständnis davon, wie menschliche Aufsicht von KI theoretisch gefasst, praktisch umgesetzt, technisch implementiert und empirisch bewertet werden kann.

Besonders deutlich wird die Herausforderung in der medizinischen Diagnostik, wenn KI-Systeme Befunde vorschlagen: Zwar verspricht deren Einsatz eine höhere Effizienz, doch es besteht die Gefahr, dass beispielsweise bestimmte Bevölkerungsgruppen systematisch benachteiligt werden – etwa durch verzerrte Trainingsdaten – und dass Ärztinnen und Ärzte sich zu sehr auf automatisierte Vorschläge verlassen. Gerade bei unauffälligen Befunden wird eine tiefere selbstständige Beschäftigung mit Patienten und vorliegenden Daten womöglich vernachlässigt.

Oberflächlich scheint hier menschliche Aufsicht gewährleistet: Schließlich trifft letztlich eine Ärztin oder ein Arzt die Diagnose. Doch das greift zu kurz. Entscheidend ist, ob diese Person tatsächlich bereit und in der Lage ist, die KI-Ergebnisse kritisch zu prüfen, Risiken zu erkennen und bei Bedarf den zusätzlichen Aufwand auf sich zu nehmen, korrigierend einzugreifen. In der ärztlichen Praxis fällt die Rolle der medizinischen Fachkraft oft mit der der KI-Aufsichtsperson zusammen.

In anderen Hochrisikokontexten sind Beteiligung und Aufsicht klar getrennt. So etwa bei vollständig autonomen Robotaxis, die in Städten wie Austin oder San Francisco bereits für den Personentransport eingesetzt werden. Hier trifft niemand im Fahrzeug Entscheidungen – die Verantwortung für Eingriffe der Aufsicht liegt bei einer externen Leitstelle mit entsprechend geschultem Aufsichtspersonal. Diese muss in der Lage sein, über geeignete Schnittstellen mehrere Fahrzeuge parallel zu überwachen und bei Bedarf einzugreifen – beispielsweise bei

Leibniz
-Gemeinschaft

Vorsitzender des Aufsichtsrates
Prof. Dr. Stefan Jähnichen
Geschäftsführung:
Prof. Dr.-Ing. Holger Hermanns
Heike Meißner

Amtsgericht Saarbrücken HRB 63800
Ust.-IdNr. DE 137972446
Bankverbindung: Bank 1 Saar
IBAN: DE39 5919 0000 0005 1000 03
BIC: SABADE5S

Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH
Oktavie-Allee
66687 Wadern, Deutschland
www.dagstuhl.de, presse@dagstuhl.de
Fon: +49 6871 905 203, Fax: +49 6871 905 220

unerwarteten Ereignissen oder Systemausfällen. Die Herausforderung liegt hier nicht nur in der technischen Realisierbarkeit und Unterstützung, sondern auch in der Gestaltung der institutionellen, rechtlichen und kognitiven Voraussetzungen für eine tatsächlich wirksame Aufsicht.

Beide Beispiele zeigen: Allein die Tatsache, dass Menschen in einem Prozess eine entscheidende Rolle spielen, genügt nicht, um menschliche Aufsicht zu gewährleisten. Damit Menschen auch wirklich Risiken abwenden, Fehlentwicklungen korrigieren oder Schaden verhindern können – und nicht bloß Teil einer Fehlentscheidung oder Sündenböcke werden –, muss ihre Rolle gezielt und wirksam gestaltet sein. Wirksame menschliche Aufsicht ist eine spezifische Funktion, die eigene Anforderungen zu erfüllen hat. Einige der im Rahmen des Seminars intensiv diskutierten zentralen Fragestellungen waren:

- Wie unterscheidet sich Aufsicht über KI von anderen Formen der Interaktion mit KI?
- Wie können technische Ansätze mit normativen Anforderungen aus Recht und Ethik zusammengebracht werden?
- Wie müssen Mensch-Maschine-Schnittstellen gestaltet sein, damit Menschen KI-Systeme effektiv überwachen und kontrollieren können?
- Welche Voraussetzungen sind auf individueller, technischer und institutioneller Ebene nötig?
- An welchen Stellen greift die EU-KI-Verordnung zu kurz?

Ziel des Seminars war es, disziplinübergreifende Perspektiven zu bündeln und die Voraussetzungen, Herausforderungen und Erfolgsfaktoren menschlicher Aufsicht über KI gemeinsam zu schärfen – als Grundlage für technische Innovation und regulatorische Wirksamkeit.

Zentrale Ergebnisse

Menschliche Mitgestaltung und Intervention sind in vielen Bereichen nötig – von der Systemwartung bis zu regulatorischen Entscheidungen. „Menschliche Aufsicht“ bezeichnet jedoch eine spezifische Form dieser Eingriffe, welche die Teilnehmer des Seminars als System definierten bei dem

- eine natürliche Person (oder mehrere natürliche Personen)
- systematisch vorbereitet die Möglichkeit hat,
- bewusst den Betrieb zu überwachen und
- bei Bedarf einzugreifen,
- um die KI-induzierten Risiken substantiell zu vermindern.

Dabei können mehrere Ebenen der Aufsicht parallel existieren: Verschiedene Personen können dasselbe System zu gleichen oder unterschiedlichen Zeiten, mit verschiedenen Schwerpunkten oder aus unterschiedlichen (auch zeitlichen) Blickwinkeln überwachen. Insbesondere wurde herausgearbeitet, dass diese Definition konkrete Konsequenzen für Design und Entwicklung von KI-Systemen hat.

Wichtig ist: Der Eingriff einer Aufsicht muss sich dabei keineswegs auf einzelne KI-Entscheidungen beschränken. Es können auch die Systemvoraussetzungen verbessert oder die zugrundeliegenden Entscheidungsprozesse optimiert werden. Menschliche Aufsicht wirkt damit sowohl operativ im laufenden Betrieb als auch strategisch auf der Systemebene.

Die Herausforderung der Restrisiken

Menschliche Aufsicht ist eine von mehreren Anforderungen der KI-Verordnung für Hochrisiko-KI-Systeme. Sie fungiert als Auffangnetz: Risiken, die sich technisch nicht vollständig ausschließen lassen, sollen vom Menschen kontrolliert und minimiert werden.

Das bedeutet: In Europa werden KI-Systeme mit beachtlichen Restrisiken auf den Markt kommen. Deren Beherrschung hängt von den technischen Möglichkeiten, den individuellen Fähigkeiten und der Motivation der Aufsichtspersonen sowie den konkreten Arbeitsbedingungen ab. Diese Erkenntnis unterstreicht die Bedeutung interdisziplinärer Forschung zu den Erfolgsfaktoren effektiver Aufsicht.

Das Seminar identifizierte dabei drei zentrale Bereiche für wirksame menschliche Aufsicht: **technische Faktoren** (wie Systemdesign, Erklärbarkeitsmethoden und Benutzeroberflächen), **individuelle Faktoren** (wie Fachkompetenz, Motivation und psychologische Aspekte der Aufsichtsperson) sowie **umgebungsbedingte Faktoren** (wie Arbeitsplatzgestaltung und organisatorische Rahmenbedingungen).

Identifizierte Herausforderungen

Das Seminar identifizierte mehrere zentrale Problemfelder, denen sich zukünftige Forschung widmen sollte:

Verantwortbarkeit: Wie verhindert man, dass der Mensch zum reinen Feigenblatt verkommt, zum willigen Erfüllungsgehilfen der KI wird oder sich rein ökonomischen Interessen beugt?

Risikomanagement: Wie stehen verschiedene Ansätze von Risikomanagement zueinander? Welche Rolle nimmt die menschliche Aufsicht darin jeweils ein?

Technische und organisatorische Unterstützung: Wie unterstützt man Personen dabei zu erkennen, wann sie eingreifen sollen – und wann nicht? Wie verhindert man, dass die Aufsichtsperson mehr Risiken schafft als sie mindert?

Kognitive Verzerrungen: Wie verhindert man, dass Menschen der Neigung nachgeben, KI-Entscheidungen unkritisch zu akzeptieren, und trotzdem effizient entscheiden?

Erfolgsmessung: Welche Maßstäbe gelten für die Effektivität menschlicher Aufsicht, insbesondere beim Schutz von Grundrechten?

Die Teilnehmer des Seminars waren sich einig: Effektive menschliche Aufsicht ist möglich, aber keineswegs trivial. Sie erfordert mehr als gute Absichten oder das Hinzufügen menschlicher Beteiligung an beliebiger Stelle im Entscheidungsprozess. Stattdessen werden systematische Verfahren zur Entwicklung von KI-Systemen, zusätzliche Werkzeuge, interdisziplinäre Zusammenarbeit, und konkrete Umsetzungsstrategien benötigt.

Das Dagstuhl-Seminar fand vom 29. Juni bis zum 4. Juli 2025 statt und brachte führende internationale Experten aus 11 Ländern von 4 Kontinenten zusammen, um die Grundlagen für eine wirksame menschliche KI-Aufsicht zu erarbeiten. Wesentliche Expertise wurde bereitgestellt von überregionalen Forschungsinitiativen, darunter der transregionale Sonderforschungsbereich 248 „Center for Perspicuous Computing“ (CPEC) der DFG sowie dem „Center for European Research for Trusted AI” (CERTAIN), einer Initiative des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI). Ansprechpartner sind hier Raimund Dachselt (CPEC, TU Dresden) und Kevin Baum (CERTAIN, DFKI).

Hintergrund:
Schloss Dagstuhl lädt das ganze Jahr über Wissenschaftler aus aller Welt ins nördliche Saarland ein um über neueste Forschungsergebnisse in der Informatik zu diskutieren. Mehr als 3.500 Informatiker von Hochschulen, Forschungseinrichtungen und aus der Industrie nehmen jährlich an den wissenschaftlichen Veranstaltungen in Dagstuhl teil. Darüber hinaus betreibt Schloss Dagstuhl mit dblp eine zentrale Publikationsdatenbank für Informatik und bietet des Weiteren auch Verlagsdienstleistungen im Bereich Open Access Publishing an. Seit 2005 gehört Schloss Dagstuhl zur Leibniz-Gemeinschaft, in der zurzeit 96 führende außeruniversitäre Forschungsinstitute und wissenschaftliche Infrastruktureinrichtungen in Deutschland vertreten sind. Aufgrund ihrer gesamtstaatlichen Bedeutung fördern Bund und Länder die Institute der Leibniz-Gemeinschaft gemeinsam.

Kontakte:

- *Prof. Dr. Markus Langer <markus.langer@psychologie.uni-freiburg.de>, Universität Freiburg, für wissenschaftliche Fragen zum Seminar*
- *Dr. Johann Laux <johann.laux@oii.ox.ac.uk>, Universität Oxford, für spezifische Presse-Anfragen*
- *Prof. Dr. Holger Hermanns <holger.hermanns@dagstuhl.de> für Schloss Dagstuhl – LZI*