SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Wadern, July 9, 2025

Schloss Dagstuhl:

Oversight of AI, can it work?

Interdisciplinary seminar on effective human oversight of AI systems at Schloss Dagstuhl

Artificial intelligence (AI) is increasingly making decisions that have a profound impact on our lives – from medical diagnostics to autonomous mobility. Hoping that everything will always go well is nothing but naive. People may get harmed or their rights be violated by AI systems suggesting incorrect diagnoses or making discriminatory decisions. It is impossible to control these risks by technical means alone. Effective human oversight is therefore a key component of the trustworthy use of AI, and is mandatory for particularly high-risk systems under the European AI Act. The Act entered into force in August 2024 and most of its rules will become fully applicable by August 2, 2026. Thus, the time is now to explore the question: What are the conditions for effective human oversight of AI?

This topic was addressed at the Dagstuhl Seminar "**Challenges of Human Oversight: Achieving Human Control of AI-Based Systems**", which took place from **June 29 to July 4, 2025**, at Schloss Dagstuhl. International experts from the fields of computer science, psychology, law, ethics, cognitive science, and technology design participated.

The workshop focused on questions surrounding the effectiveness, fairness, efficiency, and efficacy of human oversight. For example, there was intense discussion about how to avoid situations in which it is much easier to go along with a decision suggested by AI than to invest the effort needed to raise a well-founded objection against that decision.

From requirement to implementation

Terms such as "human-in-the-loop" and "human oversight" are now firmly established in political discourse. However, there is still no consistent, interdisciplinary understanding of how human oversight of AI can be conceptualized, implemented in practice, implemented technically, and evaluated empirically.

The challenge is particularly evident in medical diagnostics when AI systems suggest findings: Although their use promises greater efficiency, there is a risk that certain population groups will be systematically disadvantaged – for example, through biased training data – and that doctors will rely too heavily on automated suggestions. Particularly in the case of unspecific findings, a deeper independent examination of patients and the available data may get neglected.

Superficially, human oversight seems to be guaranteed here: after all, it is ultimately a doctor who makes the diagnosis. But that is not enough. The decisive factor is whether this person is actually willing and able to critically review the AI results, identify risks, and, if necessary, take on the additional effort of corrective action. In medical practice, the role of the medical professional often coincides with that of the AI supervisor.

In other high-risk contexts, participation and supervision are clearly separated. This is the case, for example, with fully autonomous robotaxis, which are already today transporting passengers in cities such as San Francisco. Here, no one in the vehicle takes any decisions – the responsibility for oversight intervention lies with an external control center with appropriately trained oversight staff. This center must be able to monitor several vehicles simultaneously via suitable interfaces and intervene if necessary – for example, if there are unexpected events or system failures. The challenge here lies not only in technical feasibility and support, but also in designing the institutional, legal, and cognitive prerequisites for truly effective oversight.

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

Both examples show that the mere fact that humans play a decisive role in a process is not sufficient to ensure human oversight. For humans to actually avert risks, to correct undesirable developments, or to prevent damage – instead of becoming scapegoats or part of a wrong decision – their role must be specifically and effectively designed. Effective human oversight is a separate function that has its own requirements to fulfill. Some of the key questions discussed in depth during the seminar were:

- How does oversight of AI differ from other forms of interaction with AI?
- How can technical approaches be reconciled with normative requirements from law and ethics?
- How must human-machine interfaces be designed so that humans can effectively monitor and control AI systems?
- What prerequisites are necessary at the individual, technical, and institutional levels?
- In what aspects does the EU AI Act fall short?

The aim of the seminar has been to bring together interdisciplinary perspectives and jointly sharpen the prerequisites, challenges, and success factors of human oversight of AI – as a basis for technical innovation and regulatory effectiveness.

Key findings

Human involvement and intervention are necessary in many areas – from system maintenance to regulatory decisions. However, "human oversight" refers to a specific form of intervention, which the seminar participants defined as a system in which

- a natural person (or several natural persons)
- who is systematically prepared for and is in the position to
- consciously monitor operations and
- intervene, if necessary,
- in order to substantially reduce AI-induced risks.

Several levels of oversight can exist in parallel: different people can monitor the same system at the same or different times, with different foci or from different (including temporal) perspectives. In particular, the seminar participants have highlighted that this definition has concrete consequences for the design and development of AI systems.

It is important to note that oversight intervention is by no means limited to individual AI decisions. System preconditions can also be improved or the underlying decision-making processes be optimized. Human oversight thus has both an operational effect during ongoing operations and a strategic effect at the system level.

The challenge of residual risks

Human oversight is one of several requirements of the AI Act for high-risk AI systems. It acts as a safety net: risks that cannot be completely eliminated by technical means should be controlled and minimized by humans.

This implies that AI systems with considerable residual risks will come onto the market in Europe. Controlling these risks depends on the technical capabilities, individual skills, and motivation of the oversight personnel, as well as the specific working conditions. This insight underscores the importance of interdisciplinary research on the success factors of effective oversight.

The seminar has identified three key areas for effective human oversight: **technical factors** (such as system design, explainability methods, and user interfaces), **individual factors** (such as professional competence, motivation, and psychological aspects of the supervisor), and **environmental factors** (such as workplace design and organizational conditions).

Challenges identified

The seminar identified several key problem areas that future research should address:

Accountability: How can we prevent humans from becoming mere figureheads, willing accomplices of AI, or bowing to purely economic interests?

Risk management: How do different approaches to risk management relate to each other? What role does human oversight play in each case?

Technical and organizational support: How can we help people recognize when they should intervene – and when they should not? How can we prevent oversight from creating more risks than it mitigates?

Cognitive biases: How can we prevent people from giving in to the tendency to accept AI decisions uncritically while still making efficient decisions?

Measuring success: What standards apply to the effectiveness of human oversight, especially when it comes to protecting fundamental rights?

The seminar participants agreed that effective human oversight is possible, but by no means trivial. It requires more than good intentions or adding human involvement at random points in the decision-making process. Instead, systematic procedures for developing AI systems, additional tools, interdisciplinary cooperation, and concrete implementation strategies are needed.

The Dagstuhl Seminar took place from June 29 to July 4, 2025, and brought together leading international experts from 11 countries on 4 continents to lay the foundations for effective human oversight of AI. Significant expertise was provided by the transregional Collaborative Research Center 248 "Center for Perspicuous Computing" (CPEC) of the DFG and the "Center for European Research for Trusted AI" (CERTAIN), an initiative of the German Research Center for Artificial Intelligence (DFKI). The contact persons here are Raimund Dachselt (CPEC, TU Dresden) and Kevin Baum (CERTAIN, DFKI).

Background:
During the whole year, Schloss Dagstuhl invites scientists from all over the world to come to northern Saarland in the south west of Germany to debate the newest scientific findings in informatics. More than 3,500 computer scientists from universities, research institutions, and industry take part in various scientific events at Dagstuhl each year. In addition, Schloss Dagstuhl operates dblp, a central publication database for computer science, and also offers open access publishing services. Since 2005, Schloss Dagstuhl is a member of the Leibniz Association, which connects 96 leading non-university research institutes and scientific infrastructure facilities all over Germany. Because of their national importance, the federal government and the state governments jointly fund the institutes of the Leibniz Association.

Contact:

- *Prof. Dr. Markus Langer <markus.langer@psychologie.uni-freiburg.de>, University of Freiburg, for scientific questions about the seminar*
- *Dr. Johann Laux <johann.laux@oii.ox.ac.uk>, University of Oxford, for specific press inquiries*
- *Prof. Dr. Holger Hermanns <holger.hermanns@dagstuhl.de> for Schloss Dagstuhl – LZI*

If you would like to get future press releases via email, please subscribe to our press distribution list by sending an email to presseverteiler-subscribe@rhea.dagstuhl.de.