Report from Dagstuhl Seminar 25272

# Challenges of Human Oversight: Achieving Human Control of AI-Based Systems

**Markus Langer*[1], Raimund Dachselt*[2], Q. Vera Liao*[3], Tim Miller*[4], and Nava Tintarev*[5]**

1 **Universität Freiburg, DE.** markus.langer@psychologie.uni-freiburg.de
2 **TU Dresden, DE.** raimund.dachselt@tu-dresden.de
3 **Microsoft – Montréal, CA.** veraliao@umich.edu
4 **University of Queensland – Brisbane, AU.** timothy.miller@uq.edu.au
5 **Maastricht University, NL.** n.tintarev@maastrichtuniversity.nl

**Abstract**

Human oversight is a key safeguard for AI systems, intended to mitigate risks by adding a human layer of safety and control. Oversight personnel should, for example, detect malfunctions or violations of fundamental rights such as discriminatory decision-making and intervene accordingly. Human oversight is also central to AI governance and ethics, and is mandated by Articles 14 and 26 of the EU AI Act for high-risk AI. This Dagstuhl Seminar brought together experts from artificial intelligence, human-computer interaction, human factors and psychology, philosophy and ethics, and law to explore conceptual, technical, legal, and practical dimensions of human oversight of AI. Across the seminar, participants provided perspective talks from the different disciplines and engaged in working groups and use-case specific discussions in order to establish a science of human oversight of AI systems. The main outcome of this seminar is a general framework that outlines the architecture, processes, and sociotechnical design dimensions of human oversight of AI systems.

**Seminar** June 29 – July 4, 2025 – https://www.dagstuhl.de/25272
**2012 ACM Subject Classification** Human-centered computing → HCI design and evaluation methods; Human-centered computing → HCI theory, concepts and models
**Keywords and phrases** artifical intelligence, explainable ai, human oversight, norms and regulations, safety
**Digital Object Identifier** 10.4230/DagRep.15.6.189

## 1 Executive Summary

*Markus Langer (Universität Freiburg, DE)*
*Raimund Dachselt (TU Dresden, DE)*
*Q. Vera Liao (Microsoft – Montréal, CA)*
*Tim Miller (University of Queensland – Brisbane, AU)*
*Nava Tintarev (Maastricht University, NL)*



**What is effective human oversight of AI systems?** The Dagstuhl Seminar 25272 "Challenges of Human Oversight: Achieving Human Control of AI-Based Systems" brought together interdisciplinary experts from artificial intelligence, human-computer interaction, human factors and psychology, philosophy and ethics, as well as law to explore conceptual, technical,

* Editor / Organizer


Challenges of Human Oversight: Achieving Human Control of AI-Based Systems, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 189–204
Editors: Markus Langer, Raimund Dachselt, Q. Vera Liao, Tim Miller, and Nava Tintarev
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

legal, and practical dimensions of human oversight of AI. Across the seminar, participants provided perspective talks from the different disciplines and engaged in working groups and use-case specific discussions in order to establish a science of human oversight of AI systems. The main outcome of this seminar is a general framework that outlines the architecture, processes, and sociotechnical design dimensions of human oversight of AI systems. In the following, we present some of the key insights of this seminar in more detail.

**Conceptual Foundations of Human Oversight.** Human oversight is defined as a human activity to monitor and intervene in AI-supported tasks (typically at runtime) with the aim of sufficiently mitigating risks. Mitigating risks means detecting errors, system malfunctions, or inadequate outputs. Effectiveness depends on epistemic access, causal power, self-control, and fitting intentions of the human oversight personnel. In order to optimize the human oversight effectiveness, it requires designing the sociotechnical dimensions of human oversight: the human factors, technical design, and contextual considerations. Human oversight can operate at multiple layers and across distributed roles within human oversight teams and is inherently interdependent with other risk mitigation measures.

**Human Factors, Technical Design, and Contextual Considerations.** Human factors cover situation awareness, decision making, cognitive biases, workload, motivation, training, and collaboration between oversight personnel. Technical design must support human detection of system errors and failures, for example via visualization, technical support tools, adaptive automation, handover design, and personalization. Contextual considerations to support human oversight include time and resource requirements for effective human oversight as well as the clarity of human oversight roles and duties.

**Legal, and Normative Considerations.** Human oversight effectiveness requires normative judgment beyond legal mandates. Human oversight objects include individual AI systems, highly-autonomous agentic AI in high-risk domains, as well as AI systems operated by users such as patients using mental health chatbots who themselves are not considered human oversight personnel. The seminar highlighted the importance to consider the relation between human oversight, other risk management measures, and technical standards.

**Evaluating Human Oversight Effectiveness.** Evaluation of human oversight implementation is crucial given that human oversight effectiveness can only be achieved iteratively. Metrics include effectiveness of monitoring and interventions (e.g., detecting erroneous AI outputs, overriding these outputs), alignment with human oversight protocols, and long-term performance outcomes. Mixed-methods approaches (quantitative and qualitative) and comparative studies of different human oversight design options (e.g., varying support interfaces) were discussed as possible options to evaluate human oversight effectiveness.

**Human Oversight Effectiveness as an Iterative and Multi-Layered Challenge.** Continuous updating of human oversight design is essential, integrating empirical feedback and ensuring institutional support for high-quality and sustainable human oversight of AI. Furthermore, we saw that effective oversight required identifying information and workflows across regulatory, technical, and interface layers.

**Conclusion.** The seminar demonstrated that human oversight of AI is a multifaceted, interdisciplinary challenge, involving conceptual clarity, human factors, technical design, contextual considerations, evaluation frameworks, as well as legal and ethical considerations. The outputs of this seminar provide a foundation for theoretical modeling, empirical research, practical design guidance, and normative reflection, establishing a roadmap for advancing

effective human oversight in AI systems contributing to the safe implementation of AI in high-risk contexts. Next steps include joint publications (e.g., a framework for human oversight of AI), developing technical support tools for effective human oversight, and community building through workshops at key human-computer interaction and AI conferences.

## 2 Table of Contents

**Executive Summary**
*Markus Langer, Raimund Dachselt, Q. Vera Liao, Tim Miller, and Nava Tintarev* 189

**Overview of Talks**

What is Human Oversight?
*Markus Langer and Sarah Sterz* . . . 194

Human Oversight of AI Systems: An HCI Perspective
*Ujwal Gadiraju* . . . 194

A Legal Perspective on Human Oversight
*Anne Lauber-Rönsberg* . . . 195

Towards Human Oversight of Imperfect Automation: A Dagstuhl CELLAR (Cognitive Engineering Lessons Learned And Reflections) Perspective
*Tim Miller and Liz Sonenberg* . . . 196

From Traditional Auditing to Everyday Oversight: The Role of Users in Algorithmic Accountability
*Motahhare Eslami* . . . 196

**Working groups**

Thematic Analysis of Lightning Talks
*Raimund Dachselt, Susanne Gaube, Tim Miller, Liz Sonenberg, and Nava Tintarev* 197

Black Mirror Writers' Room Exercise
*Nava Tintarev* . . . 197

Use Case Human Resource Management
*Anna Maria Feit, Harmanpreet Kaur, Mark T. Keane, Richard Landers, Markus Langer, and Q. Vera Liao* . . . 198

Use Case Autonomous Driving
*Oana Inel, Linda Onnasch, and Carola Plesch* . . . 199

Defining and Conceptualizing Human Oversight
*Kevin Baum, Markus Langer, Anne Lauber-Rönsberg, Johann Laux, Tim Schrills, and Sarah Sterz* . . . 199

Dimensions of Human Oversight of AI
*Virginia Dignum, Ujwal Gadiraju, Brian Lim, Marija Slavkovik, Chenhao Tan, Ziang Xiao, and Hanwei Zhang* . . . 200

Challenges of Human Oversight – Human Factors Challenges
*Anna Maria Feit, Liz Sonenberg, Markus Langer, Q. Vera Liao, and Linda Onnasch* 200

Challenges of Human Oversight – Technical Challenges
*Brian Lim, Chenhao Tan, Ziang Xiao, and Hanwei Zhang* . . . 201

Challenges of Human Oversight – Legal Challenges
*Anne Lauber-Rönsberg, Johann Laux, Philip Meinel, and Silja Voeneky* . . . 201

Challenges of Human Oversight – Evaluation Challenges
*Raimund Dachselt, Susanne Gaube, Holger Hermanns, Oana Inel, Mark T. Keane, Tim Miller, Carola Plesch, and Nava Tintarev* . . . 202

**Integration and Outlook**

Bringing it all Together
*Raimund Dachselt, Markus Langer, Q. Vera Liao, Tim Miller, and Nava Tintarev* 202

Future Activities
*Raimund Dachselt, Markus Langer, Q. Vera Liao, Tim Miller, and Nava Tintarev* 203

**Participants** . . . . . 204

# 3 Overview of Talks

Before the seminar, we prompted input on the topic of human oversight of AI from the different disciplines that were part of the seminar. We selected five talks to be presented during Monday and Tuesday of the seminar that provided a starting point for interdisciplinary discussions on human oversight of AI.

## 3.1 What is Human Oversight?

*Markus Langer (Universität Freiburg, DE) and Sarah Sterz (Universität des Saarlandes – Saarbrücken, DE)*


**Joint work of** Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Markus Langer
**Main reference** Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, Markus Langer: "On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives", in Proc. of the The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024, pp. 2495–2507, ACM, 2024.
**URL** https://doi.org//10.1145/3630106.3659051

Human oversight is discussed as a potential safeguard to mitigate risks in AI applications. This prompts a critical examination of the role and conditions necessary for effective or meaningful human oversight of these systems. Based on the claim that the main objective of human oversight is risk mitigation, we propose a viable understanding of effectiveness in human oversight: for human oversight to be effective, the oversight person has to have (a) sufficient control and power with regard to the system and its effects, (b) suitable epistemic access to relevant aspects of the situation, and (c) fitting intentions for their role. Furthermore, we argue that this is equivalent to saying that an oversight person is effective if and only if they are morally responsible and have fitting intentions. Against this backdrop, we suggest facilitators and inhibitors of effectiveness in human oversight when striving for practical applicability. We discuss factors in three domains, namely, the technical design of the system, individual factors of oversight persons, and the environmental circumstances in which they operate.

## 3.2 Human Oversight of AI Systems: An HCI Perspective

*Ujwal Gadiraju (TU Delft, NL)*



The systematic involvement of humans in monitoring, controlling, and intervening in AI system operations to ensure safety, accountability, and alignment with human values has a central role to play in regulating meaningful AI adoption. This talk synthesizes challenges and opportunities for human oversight of AI systems from a human-computer interaction (HCI) standpoint by addressing the overarching goal of designing and developing interfaces, interaction paradigms, and workflows that enable effective human-AI collaboration while maintaining meaningful human control. This perspective talk concludes with a discussion around the critical components of human oversight, and potential methods and measures for effective human oversight.

## 3.3 A Legal Perspective on Human Oversight

*Anne Lauber-Rönsberg (TU Dresden, DE)*



Objectives that can be pursued by human oversight are risk mitigation, trustworthiness of AI systems, human agency and autonomy, human-centered AI and accountability. There are many facets of human involvement that contribute to these objectives: human approval of AI output (human in the loop); monitoring of AI decision processes and the possibility to intervene (human on the loop); reducing the role of the AI-system to supporting human decisions; human interventions in the operation of the AI system, such as correction of inaccurate results or the infamous stop button and finally a human examination of AI-output after contestation. These activities are either executed during run time, maybe even real time, during inspection time and supported by decisions on the technical design during design time. Monitoring activities can either relate to every output or to selected samples. They may be always required, for instance by the law, or only be triggered upon the request of an affected person. I highlight that the concept of human oversight of Art. 14 and 26 of the EU AI Act is much more confined since it relates only to a subset of the activities named. Human oversight for high-risk AI systems under the AI Act aims at mitigating risks to health, safety or fundamental rights and must be be commensurate with the risks, level of autonomy and context of use, has to take place during the use of the system and requires the ability of the oversight person to intervene in the operation of the AI system, such as by stopping it or by overriding its output. Apart from this, the AI Act contains few specifications regarding technical design. Thus, the main responsibility is on the provider (= developer), who has to determine appropriate oversight measures that are either built into the system (such as human-machine interface tools or operational constraints) or have to be executed by the deployer in line with the instructions for use. The deployer has to assign human oversight to persons who have the necessary competence, training and authority and needs to ensure the necessary support. Oversight persons can be employees who use the AI system for their work or external third parties. End-users in a private context are not required to conduct human oversight by the AI Act. Questions to be discussed are: Under which conditions and in which contexts can human oversight be regarded as (sufficiently) effective? How can technical standards be drafted to clarify the legal obligations? How can human oversight be implemented in case of more autonomous agentic AI systems? What applies in situations where there is no deployer (e.g., in case of smart toys)? The legal obligations to provide human oversight are not applicable to AI systems that are only "intended to perform a preparatory task". How is this exception to be construed?

## 3.4 Towards Human Oversight of Imperfect Automation: A Dagstuhl CELLAR (Cognitive Engineering Lessons Learned And Reflections) Perspective

*Tim Miller (University of Queensland – Brisbane, AU) and Liz Sonenberg (University of Melbourne, AU)*



In this perspective talk, we profile a selection of research from cognitive science and cognitive systems engineering that has high relevance to human oversight in situations involving artificial intelligence. We look back at research on human interaction with imperfect automation, that captures automation-induced failures that can occur at a technical level, through the human-automation interaction, and/or through broader sociotechnical systems breakdowns. We show there is a rich existing literature that can inform contemporary analyses of human oversight of AI systems, and illustrate that many of the "new" problems in human oversight of AI are not so new at all. We also present a more future-oriented view of emerging paradigms in cognitive science/psychology, and how these may affect approaches to characterizing and designing the AI and the great sociotechnical system in which it operates.

## 3.5 From Traditional Auditing to Everyday Oversight: The Role of Users in Algorithmic Accountability

*Motahhare Eslami (Carnegie Mellon University - Pittsburgh, US)*



This talk examines a growing shift in algorithmic accountability: from formal, expert-led audits to more informal, everyday forms of oversight performed by users themselves. While traditional audits remain essential, they often fail to capture the full range of harms experienced by the public – particularly those affecting marginalized communities. In contrast, users are increasingly taking initiative to investigate, document, and call out algorithmic failures, often through grassroots experimentation and collaborative sensemaking.

Through a series of empirical studies, the talk traces how users have identified and investigated algorithmic harms across platforms. These forms of "everyday oversight" not only surface harms often missed by formal audits but also broaden the notion of who gets to hold AI systems accountable. To support this growing public role, the talk introduces WeAudit, a platform co-designed with users and industry practitioners to facilitate participatory auditing at scale. By recognizing users as legitimate and capable auditors, WeAudit aims to institutionalize support for community-driven accountability efforts.

# 4 Working groups

From Tuesday to Thursday, main activities were organized in working groups where we worked on conceptual and design challenges of human oversight of AI. In the following, we provide summaries of the working groups. The order of authors in the working groups is alphabetical.

## 4.1 Thematic Analysis of Lightning Talks

*Raimund Dachselt (TU Dresden, DE), Susanne Gaube (University College London, GB), Tim Miller (University of Queensland – Brisbane, AU), Liz Sonenberg (University of Melbourne, AU), and Nava Tintarev (Maastricht University, NL)*



This working groups conducted a thematic analysis of the lightning talks, where participants introduced their research and discussion interests during the first day of the seminar. The analysis surfaced a rich landscape of perspectives on human oversight, which clustered into four broad thematic areas. First, participants highlighted applications and use cases, supported by a taxonomy of types of human oversight and intervention, and situated oversight in relation to adjacent concepts such as accountability and responsibility. Second, a legal, risk, and ethics theme emerged, including references to the EU AI Act, legal processes, risk management practices, and the need for future-proof regulatory frameworks. Third, participants stressed the role of human factors, from cognitive limitations and strengths to capability development, user interaction design, and explainable AI. Finally, a set of themes focused on design, evaluation, and technical development, encompassing design methodologies, information visualization, evaluation and measurement approaches, traceability, modeling, and the technical capabilities of AI systems. Together, these clusters reflect the interdisciplinary scope of the seminar and lay a foundation for deeper discussions on human oversight.

## 4.2 Black Mirror Writers' Room Exercise

*Nava Tintarev (Maastricht University, NL)*



Authors for this working group are all participants. Nava Tintarev moderated the working group.

Black Mirror is science-fiction series that offers a critical discussion and implications of technological developments that are expected in the near future (5 years). Each episode focuses on a specific technology, e.g., social credit scores in the episode "Nosedive", inspired by the social credit system developed in China. All of these episodes are dystopian, looking at the dark side of technology (dystopian means relating to or denoting an imagined state or society where there is great suffering or injustice). The task for participants was to write their own Black Mirror scenario involving a specific task context relevant to the topic of human oversight of AI. These are the teasers of the Black Mirror episodes pitched by the participant groups.

**Train Wreck.** An employee of the government transport department notices massive chaos in public transport. When questioning his superiors, he finds out that the cause is decisions made by his Gov3.0 productive tool, that he was incentivized to use for efficiency reasons... and that he authorized himself.

**Border Control AI.** Otto once found deep meaning in his work as a border agent, making autonomous decisions that impacted national security – but as AI systems take over, his role becomes reduced to a mere formality. Despite new oversight measures meant to ensure human responsibility, Otto feels powerless, resorting to workarounds while struggling with burnout, dwindling job prospects, and a growing sense of despair.

**AI Teacher.** A suicide epidemic sweeps the country. Unemployment is high and there is an overabundance of unhappy bus drivers. In the meantime care jobs are left unfilled. Jeremy the high-school student wants to study nursing, but his personalized AI tutor only teaches traffic regulations and driving simulations. His parents ask the AI tutor for information about other personalized curricula. While the parents do not get any answers, this starts a positive cascade to the district educational office and the AI allocation module. Follow these parents in their brave battle against bureaucratic demons.

**Professor Bot.** A university introduces a new software for professors to create digital copies of themselves to "scale" advising. Professor Best had the best intention by creating a "better" version of his digital copy ... Who will stay? Professor Best or ProfessorBot?

**BlueSky.** The world is in an energy crisis. We have exhausted oil and making no progress in fusion. The governments have opened a Blue Sky grant call for groundbreaking ideas on how to solve the energy crises. Generation one of grants has offered no impact. A generation two call is out. Alex is considering a proposal. They found a old fashioned organic book in their basement on particle physics. Alex makes a proposal but it is rejected. Forty years ago research peer review was replaced by AI. Particle physics has been systematically denied funding because it is very low practical impact and there are real problems in society.

## 4.3 Use Case Human Resource Management

*Anna Maria Feit (Universität des Saarlandes – Saarbrücken, DE), Harmanpreet Kaur (University of Minnesota – Minneapolis, US), Mark T. Keane (University College Dublin, IE), Richard Landers (University of Minnesota – Minneapolis, US), Markus Langer (Universität Freiburg, DE), and Q. Vera Liao (Microsoft – Montréal, CA)*



This working group explored human oversight in job–applicant matching platforms, framing it as a process of detecting failures and taking appropriate action. Oversight applies at two levels: monitoring algorithmic performance and reviewing individual matches. The group distinguished between aggregate risks (e.g., long-term discrimination) and individual risks. Detection of inadequate system outputs may rely on automated thresholds or manual checks, while interventions range from alerting decision makers and redoing tasks to retraining models or redesigning the system.

Effective oversight requires dashboards and feedback loops that integrate aggregate and individual data, reliable alerts, and clarity about the authority of oversight roles. Challenges include avoiding overreliance on oversight, balancing efficiency with risk mitigation, and defining what a "safe state" means in non-real-time systems. At a meta-level, oversight design must be continuously updated as new problems arise, with institutions ensuring its quality over time.

This working group also discussed human oversight of coaching AI, where a human oversight person oversees an interaction between a conversational AI and a user. This use case also inspired discussions on other high-risk conversational AI domains such as AI in psychotherapy, where human oversight may require a control room where human oversight personnel can oversee and intervene in patient-AI conversations, or AI used by children where human oversight may be performed by parents.

## 4.4 Use Case Autonomous Driving

*Oana Inel (Universität Zürich, CH), Linda Onnasch (TU Berlin, DE), and Carola Plesch (BSI – Bonn, DE)*



This working group explored human oversight in the use case of autonomous driving. For example, a remote oversight person oversees a fleet of vehicles within a defined area, but without continuous monitoring. Instead, autonomous vehicles escalate critical situations to the oversight person, while time-critical decisions remain with the vehicle itself. Passengers have no direct intervention options beyond contacting the oversight person.

The group characterized this setting as remote, multitasking oversight, with oversight tasks interrupting other activities through forced task switches. Key questions included whether a "safe state" exists for the vehicle, and whether it can be unconditionally enforced. Intervention options available to the oversight person largely concern planning decisions (e.g., route selection), escalation measures (e.g., contacting a task force), or putting a vehicle into a safe state. Interfaces must integrate multiple streams of information – traffic, weather, vehicle status – while supporting interaction with passengers and prioritization across simultaneous requests.

The group identified significant challenges for oversight effectiveness, such as establishing situation awareness from a distance, coping with limited multimodal information, managing authority boundaries between human and automation, and handling fluctuating workloads. For effective human oversight, the group emphasized the importance of contextual information provided prior to takeover, such as vehicle type, location, reason for the request, and possible courses of action, potentially offered through structured, pre-selected options.

## 4.5 Defining and Conceptualizing Human Oversight

*Kevin Baum (DFKI – Saarbrücken, DE), Markus Langer (Universität Freiburg, DE), Anne Lauber-Rönsberg (TU Dresden, DE), Johann Laux (University of Oxford, GB), Tim Schrills (Universität Lübeck, DE), and Sarah Sterz (Universität des Saarlandes – Saarbrücken, DE)*



This working group focused on defining and conceptualizing human oversight of AI systems. They proposed that human oversight consists of monitoring and intervening in AI-supported tasks with the explicit aim of mitigating risks. Oversight is effective only when these risks are sufficiently reduced. The group emphasized that effectiveness depends not only on individual abilities and motivation, but also on technical design, organizational context, and available interventions.

They identified four key factors – epistemic access, causal power, self-control and fitting intentions – that shape oversight effectiveness, and noted that different layers (legal, institutional, design/technical) influence these factors in distinct ways. Human oversight was seen as the last layer of risk mitigation, highly interdependent with other safeguards, and potentially distributed across multiple people and roles.

The group stressed that while training, design, and organizational conditions are not themselves "oversight," they are necessary enablers for it to be effective. A structured overview of oversight activities and effectiveness factors can serve as a foundation for theoretical models and guide empirical research.

## 4.6 Dimensions of Human Oversight of AI

*Virginia Dignum (University of Umeå, SE), Ujwal Gadiraju (TU Delft, NL), Brian Lim (National University of Singapore, SG), Marija Slavkovik (University of Bergen, NO), Chenhao Tan (University of Chicago, US), Ziang Xiao (Johns Hopkins University – Baltimore, US), and Hanwei Zhang (Universität des Saarlandes – Saarbrücken, DE)*



This working group outlined an dimensions of human oversight of AI. These dimensions included oversight goals, oversight tasks, oversight persons and their characteristics and skills, oversight failures and oversight evaluation. This working group also discussed the importance of proportionality of human oversight depending on the context of use and the efficient design of interfaces to support human oversight.

## 4.7 Challenges of Human Oversight – Human Factors Challenges

*Anna Maria Feit (Universität des Saarlandes – Saarbrücken, DE), Liz Sonenberg (University of Melbourne, AU), Markus Langer (Universität Freiburg, DE), Q. Vera Liao (Microsoft – Montréal, CA), and Linda Onnasch (TU Berlin, DE)*



This working group focused on the human factors and design considerations of oversight systems. For run-time oversight, they highlighted classic human factors issues such as information processing, decision making, cognitive and automation biases, uncertainty, attention management, workload, motivation, training, and collaboration. This was also described as "old human factors wine in a new bottle."

They also discussed design requirements for technical components that support human oversight: interfaces and alerts, signal processing and visualization, tools for decision support and intervention, transparency of both AI and the broader human–AI task system, history tracking, partial automation of monitoring or actions, handovers, and personalization for oversight personnel.

Finally, the group noted the dependencies between oversight systems and the task systems they oversee, and raised questions about human factors challenges for non–run-time overseers (e.g., auditors), suggesting that interface and tool design should differ across oversight roles.

## 4.8 Challenges of Human Oversight – Technical Challenges

*Brian Lim (National University of Singapore, SG), Chenhao Tan (University of Chicago, US), Ziang Xiao (Johns Hopkins University – Baltimore, US), and Hanwei Zhang (Universität des Saarlandes – Saarbrücken, DE)*



This working group investigated the technical challenges of enabling and supporting human oversight. They structured their discussion around four main dimensions. First, preparing information for oversight, including explainability, interpretability, calibration, and aggregation of explanations. Second, monitoring, with emphasis on detecting issues through structured information flows, context operationalization, robustness, adaptivity to drift, and third-party risk assessments. Third, intervention, covering controllability and steerability of AI models, identifying and operationalizing fail-safe states, fallback and recovery mechanisms, and managing hyperparameter sensitivity. Fourth, testing and evaluation, highlighting the value of sandbox environments, simulation, data synthesis, long-term evaluation, and methods for quantifying both human effort and systemic risks.

Additional themes included the need for personalization versus generalization in oversight tools, ensuring scalability, efficiency, and privacy (e.g., double-blind oversight), and recognizing impossibility results in operationalizing oversight (as with fairness definitions). The group underscored the importance of addressing these challenges in a cost-effective way while safeguarding both human and system effectiveness.

## 4.9 Challenges of Human Oversight – Legal Challenges

*Anne Lauber-Rönsberg (TU Dresden, DE), Johann Laux (University of Oxford, GB), Philip Meinel (TU Dresden, DE), and Silja Voeneky (Universität Freiburg, DE)*



This working group examined normative, legal, and organizational dimensions of effective human oversight. They noted that while effectiveness cannot be determined by legal criteria alone, any threshold of "sufficient" oversight requires normative judgment. Discussion centered on the objects of oversight, such as which AI systems and users of AI should be overseen, including highly autonomous or composite systems in high-risk domains. This raised questions about whether the EU AI Act permits or restricts highly autonomous AI in such contexts.

The group also considered the relationship between human oversight and other risk mitigation measures, identifying potential sources of failure in relation to fundamental rights and ways oversight might address them. They highlighted the importance of role distribution, including centralized versus distributed oversight, requirements for outsourcing oversight functions, and the role of deployers in supporting effective oversight. Finally, the group noted the need for technical standards to guide the design and implementation of oversight in practice.

## 4.10 Challenges of Human Oversight – Evaluation Challenges

*Raimund Dachselt (TU Dresden, DE), Susanne Gaube (University College London, GB), Holger Hermanns (Universität des Saarlandes – Saarbrücken, DE), Oana Inel (Universität Zürich, CH), Mark T. Keane (University College Dublin, IE), Tim Miller (University of Queensland – Brisbane, AU), Carola Plesch (BSI – Bonn, DE), and Nava Tintarev (Maastricht University, NL)*



This working group focused on evaluation approaches for human oversight. They emphasized the value of both comparative studies (e.g., different interface designs) and mixed methods combining quantitative with qualitative analysis of human oversight effectiveness. The group proposed metrics at multiple levels: organizational (e.g., documentation, global success indicators, costs); oversight personnel (e.g., knowledge about the domain/AI model/oversight task, motivation, efficiency, cognitive load); AI systems (e.g., performance with and without oversight); and the oversight support technology (e.g., interpretability, effectiveness in detection, predictive performance). For interventions, they suggested assessing both the effectiveness and efficiency of actions (e.g., time to resolution, coverage of delegation, alignment with protocols). Together, these dimensions offer a comprehensive framework for evaluating oversight across technical, human, and organizational layers.

# 5 Integration and Outlook

On Thursday evening and Friday, the seminar conducted integrative working group and a collective outlook session. In the following, we summarize these sessions.

## 5.1 Bringing it all Together

*Raimund Dachselt (TU Dresden, DE), Markus Langer (Universität Freiburg, DE), Q. Vera Liao (Microsoft – Montréal, CA), Tim Miller (University of Queensland – Brisbane, AU), and Nava Tintarev (Maastricht University, NL)*



All participants were part of this session. It was moderated by the organizers of the seminar.

This working group led to the development of a general human oversight architecture, including the AI task to be overseen, the human oversight task, the human oversight personnel, hierarchical layers of human oversight, and sociotechnical human oversight support design. It also led to the definition of human oversight processes and tasks, including monitoring and intervention in AI operations. Additionally, this working group synthesized challenges of human oversight in practice, such as the context-dependency and effectiveness of human oversight, challenges in evaluating human oversight, and challenges in complying with human oversight regulations.

## 5.2 Future Activities

*Raimund Dachselt (TU Dresden, DE), Markus Langer (Universität Freiburg, DE), Q. Vera Liao (Microsoft – Montréal, CA), Tim Miller (University of Queensland – Brisbane, AU), and Nava Tintarev (Maastricht University, NL)*



During the final day, the entire group discussed next steps and opportunities for collaboration. Several joint publications were proposed, including a framework paper on human oversight, a taxonomy of scenario attributes and oversight requirements, a design fiction paper based on the "Black Mirror" exercise, and follow-up work on the overlay of the EU AI Act and sustainable human oversight. Further ideas included papers on legal obligations around high-risk AI systems, dual-use risks of oversight, and a possible Dagstuhl Manifesto synthesizing outcomes.

The group also identified venues for dissemination and engagement, such as CHI, FAccT, IUI, CSCW and AI-focused conferences (NeurIPS, ICML, AAAI). Planned activities include workshops at major venues, exploration of special issues in journals (e.g., AI Magazine, Technology, Mind and Behavior), and the development of a software toolkit to make oversight testable.

Finally, the group outlined opportunities for funding and collaboration, including a COST Action proposal, EU and Australian grants, and partnerships such as the Dutch National Police project. They also encouraged exchanges, such as PhD visits across participating institutions, to sustain the momentum of the seminar.

## Participants


- Kevin Baum
DFKI – Saarbrücken, DE
- Raimund Dachselt
TU Dresden, DE
- Virginia Dignum
University of Umeå, SE
- Anna Maria Feit
Universität des Saarlandes – Saarbrücken, DE
- Ujwal Gadiraju
TU Delft, NL
- Susanne Gaube
University College London, GB
- Holger Hermanns
Universität des Saarlandes – Saarbrücken, DE
- Oana Inel
Universität Zürich, CH
- Harmanpreet Kaur
University of Minnesota – Minneapolis, US
- Mark T. Keane
University College Dublin, IE
- Richard Landers
University of Minnesota – Minneapolis, US
- Markus Langer
Universität Freiburg, DE
- Anne Lauber-Rönsberg
TU Dresden, DE
- Johann Laux
University of Oxford, GB
- Q. Vera Liao
Microsoft – Montréal, CA
- Brian Lim
National University of Singapore, SG
- Philip Meinel
TU Dresden, DE
- Tim Miller
University of Queensland – Brisbane, AU
- Linda Onnasch
TU Berlin, DE
- Carola Plesch
BSI – Bonn, DE
- Tim Schrills
Universität Lübeck, DE
- Marija Slavkovik
University of Bergen, NO
- Liz Sonenberg
University of Melbourne, AU
- Sarah Sterz
Universität des Saarlandes – Saarbrücken, DE
- Chenhao Tan
University of Chicago, US
- Nava Tintarev
Maastricht University, NL
- Silja Voeneky
Universität Freiburg, DE
- Ziang Xiao
Johns Hopkins University – Baltimore, US
- Hanwei Zhang
Universität des Saarlandes – Saarbrücken, DE