ISSN 2186-7437

NII Shonan Meeting Report

No. 239

Building Trustworthy and Interactive Recommender Systems through Argumentation

Matthias Kraus
Wolfgang Minker
Yuki Matsuda

January 19 - 22, 2026

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Introduction

As the amount of available data continues to grow at an exponential rate, it seems to be impractical for humans to manually process and review this information. Thus, recommender systems are aiding the process of decision-making and discovering new content. In today's digital landscape, recommender systems have become an integral part of the user experience on platforms such as Google, Netflix, Spotify and Amazon. This is why it is important that users have confidence in the suggestions made by these systems. To foster such trust, transparency in how recommendations are generated is essential. Therefore, users should not only understand why certain suggestions are made but also have the possibility to interact with the system to refine and argue about these recommendations. Interacting in such a way may help to build trust by challenging the system instead of blindly trusting it.

The aim of this Shonan Meeting has been to explore how recommender systems can be enhanced through argumentation and explanation capabilities, and how these enhancements impact users' trust and conversational experience. We have started by discussing current recommendation technologies, including the ways in which conversational systems facilitate users in achieving recommendation-related goals through multi-turn dialogue. Building on this foundation, we have examined contemporary approaches to argument design and explainable AI (XAI), and evaluate their effects on perceived trustworthiness, usability, and ethical considerations.

Without any doubt, there are several challenges in building trustworthy and interactive recommendation systems. The main challenge is to develop technological models, methods and strategies for designing trustworthy interactive recommendation systems using argumentation, thereby addressing several aspects:

Recommendations made by trustworthy and interactive recommender systems need to be made transparent using methods from the field of XAI. Furthermore, these explanations can be substantiated by including general information available through information retrieval and argument mining. The integration of state-of-the-art large language models further enhances the system's natural language understanding, generation and contextual reasoning capabilities. By implementing interaction based on conversation and natural language, users can intuitively communicate with the system. To this end, the dialogue and argumentation strategy applied should consider factors such as persuasion, negotiation, proactivity and awareness of social, cultural and contextual circumstances. Information and argument retrieval represent another key challenge. The use of advanced information retrieval techniques and natural language processing makes it possible to construct persuasive arguments from large data sets. Information and argument retrieval are complemented by the need for knowledge reasoning, where sophisticated reasoning skills help to understand complex relationships and make accurate recommendations. Finally, the role of multimodal input and output, including speech, gesture and emotion, cannot be overstated. Using different sensory data allows us to create more engaging and effective interactions.

An important area of discussion is how these advanced technologies affect society, end users, and industry. For industry, integrating argumentation and XAI in recommender systems potentially enhances trust, user engagement, and competitive differentiation by providing transparent, personalized, and context-aware recommendations. For society, these advancements could promote inclusivity, digital literacy, and ethical AI adoption, fostering trust in AI systems while addressing biases and accessibility. For end users, transparent and interactive recommendations may improve satisfaction by enabling users to refine suggestions, make informed decisions, and enjoy intuitive,

multimodal experiences. However, the potential negative impact on society and end users in particular needs to be discussed as well. For example, for society, poorly designed systems may perpetuate biases, raise privacy concerns due to increased data collection, and widen the digital divide by favoring resource-rich users. For end users, such systems could lead to cognitive overload, frustration with complex interactions, or a loss of autonomy if argumentation feels manipulative.

Building trustworthy and interactive recommender systems requires an active collaboration between multiple research disciplines, such as computer science, psychology, and ethics. This has resulted in a platform to exchange ideas and benefit from complementary work. The Shonan Meeting provided a unique venue for discussion and collaboration among experts from these disciplines. It was particularly valuable for identifying potential challenges and collaboratively shaping a research agenda for key directions.

Overview of the meeting

The Shonan Meeting No. 239 ("Building Trustworthy and Interactive Recommender Systems through Argumentation") took place in January 2026, from Monday 19 to Thursday 22, with 25 participating researchers from nine different countries (Japan, Germany, France, UK, USA, The Netherlands, Austria, Switzerland and Finland). The meeting was initially organized around four working groups: (a) Ethics, Manipulation and Responsible Deployment, (b) Evaluation and Impact of Argument-Based Recommendations, (c) Personalization and Interaction Design, (d) Models, Architectures and Data Foundations.

Within these groups, participants examined the current landscape of building trustworthy and interactive recommender systems through argumentation, identified key challenges, and outlined directions for future work. The discussions touched on topics of Explainability of recommendations, Trust, Persuasion and negotiation, Proactivity, Argument quality, Personalization, Social- and context-awareness, Investigation of long-term vs. short-term relations, Novel evaluation paradigms, Ethics and societal impact as well as Social responsibility.

Based on the interests indicated beforehand, participants were distributed across the four working groups. At the beginning of each day, each group presented an update on the previous day's discussions and findings to all Shonan participants. Notes from these sessions were collected in shared Overleaf slides using a common LaTeX template, forming the basis of the daily presentations. As discussions evolved, it became apparent that the topics of (d) Models, Architectures and Data Foundations were strongly intertwined with the other themes. Consequently, the initial four-group structure was consolidated into three groups to better reflect the main discussion streams and to allow for more focused and in-depth exchanges. After the first day, participants were encouraged to switch groups to contribute to multiple themes, engage with colleagues' work, and provide fresh perspectives and critical thoughts. On the final day, most participants returned to their original group to consolidate and finalize the outcomes of the group work. In addition, especially during the first two days, the meeting was complemented by twelve short "lightning talks" from senior researchers, who introduced their perspectives and research related to the seminar topic.

Meeting Schedule

The meeting followed a structured four-day program combining plenary sessions, lightning talks, and intensive working group activities (see Fig. 1).

Building Trustworthy and Interactive Recommender Systems through Argumentation					
Time Table	**Arrival Day**	**1st Day**	**2nd Day**	**3rd Day**	**Final Day**
	January 18th (Sunday)	**January 19th**	**January 20th**	**January 21st**	**January 22nd**
7:00 - 7:30					
7:30 - 8:00		Breakfast	Breakfast	Breakfast	Breakfast
8:00 - 8:30					
8:30 - 9:00					
9:00 - 9:30		Introduction	Present Results	Present Results	WG - (4 groups) - finalize
9:30 - 10:00		Interest & Expectations	4*(10min talk + 5min Q&A)	4*(10min talk + 5min Q&A)	
10:00 - 10:30		25*(2min,1 slide)			
10:30 - 11:00		Break	Break	Break	Break
11:00 - 11:30		4 Lightning Talks (Session 1) +Prof. Abraham Bernstein +Prof. Kaoru Sumi +Dr. Mark Colley (10 min talk, 5 min Q&A)	4 Lightning Talks (Session 3) + Dr. Khalid Al-Khatib + Prof. Yutaka Arakawa + Prof. Martin Baumann + Dr. John Lawrence	WG - (4 groups) - swap back to initial	Presentation of final Results
11:30 - 12:00			(10 min talk, 5 min Q&A)		
12:00 - 12:30		WG Allocation (4 groups)	Photo Shoot		Wrap up and Farewell
12:30 - 13:00		Lunch	Lunch	Lunch	Lunch
13:00 - 13:30					
13:30 - 14:00				Excursion:	
14:00 - 14:30		4 Lightning Talks (Session 2) + Prof. Keiichi Yasumoto + Prof. Elisabeth André + Dr. Vera Schmitt + Prof. David Traum	WG - (4 groups) - first swap	Visiting Kenchoji Temple (learn about "Zazen")	
14:30 - 15:00		(10 min talk, 5 min Q&A)			
15:00 - 15:30	Regular check-in	WG - (4 groups)			
15:30 - 16:00					
16:00 - 16:30		Break	Break		
16:30 - 17:00		WG - (4 groups)	WG - (4 groups) - second swap		
17:00 - 17:30					
17:30 - 18:00					
18:00 -18:30					
18:30 - 19:00		Dinner & Socializing	Dinner & Socializing		
19:00 - 19:30	Welcome Banquet			Banquet	
19:30 - 20:00					
20:00 - 20:30					
20:30 - 21:00					
21:00 - 21:30					
21:00 -22:00					

Figure 1: The schedule of the Shonan Seminar 'Building Trustworthy and Interactive Recommender Systems through Argumentation'.

On the first day (January 19), the meeting began with an introduction session, where participants briefly presented their interests and expectations. This was followed by two sessions of lightning talks by invited researchers. In the afternoon, participants were assigned to four working groups and started their initial discussions.

The second day (January 20) focused on deepening the discussions. The morning included presentations of intermediate group results, followed by an additional lightning talk session. After a photo session and lunch, participants continued their work in the working groups. A scheduled group swap in the afternoon enabled participants to contribute to different topics and engage with multiple perspectives. A second working group session took place later in the day.

On the third day (January 21), the morning was again dedicated to presenting updated group results. Participants then returned to their initial working groups to consolidate their discussions and integrate insights gained during the group exchanges. In the afternoon, a social excursion to Kenchoji Temple, including an introduction to the techniques of Zazen, was organized, followed by the banquet dinner.

The final day (January 22) was dedicated to finalizing the outcomes of the working groups. Each group presented its results in a plenary session, followed by a wrap-up discussion and farewell.

Besides the intensive and productive discussions, the Shonan Village Center offered many opportunities for networking, socializing, and relaxation, including shared lunches and dinners, a welcome reception, and various leisure facilities such as a swimming pool, table tennis, and a common lounge for evening gatherings.

List of Participants

- Matthias Kraus - LMU, Germany
- Yuki Matsuda - Okayama University, Japan
- Wolfgang Minker- Ulm University, Germany
- Patrick Gebhard- DFKI Saarbrücken, Germany
- Carolin Schindler- Ulm University, Germany
- Mark Colley - University College London, UK
- Jacqueline Urakami - Kyocera, Japan
- Elisabeth André - University of Augsburg, Germany
- Florian Pecune - University of Bordeaux, France
- Keiichi Yasumoto - NAIST, Japan
- Yutaka Arakawa - Kyushu University, Japan
- Kaoru Sumi - Hakodate Future University, Japan
- Khalid Al-Khatib - University of Groningen, The Netherlands
- John Lawrence - University of Dundee, UK
- Martin Baumann - Ulm University, Germany
- Vera Schmitt - TU Berlin / DFKI, Germany
- Kristiina Jokinen - AIST, AIRC, Japan
- Graham Wilcock - University of Helsinki, Finland
- Thomas Rist - Augsburg University of Applied Sciences, Germany
- Joel Fischer - University of Nottingham, UK
- Patricia Kahr - University of Zurich, Switzerland
- Ko Watanabe - RPTU Kaiserslautern/DFKI, Germany
- Abraham Bernstein - University of Zurich, Switzerland
- David Traum - University of Southern California, USA

List of Lightning Talks

Personalized Argumentation in Human–AI Interaction: Leveraging Emotion, Timing, and Agent Design

Kaoru Sumi - Hakodate Future University, Japan

This lightning talk discussed how personalized argumentation can contribute to building trustworthy and interactive recommender systems in human–AI interaction. The talk highlighted that the effectiveness of arguments depends not only on their content but also on who presents them, how they are expressed, and when they are delivered. Drawing on research in human–agent interaction and affective computing, examples were presented showing how agent identity, emotional expressions, and users' affective states influence the acceptance of explanations and persuasive messages. Mul-timodal sensing techniques, such as facial expressions and behavioral signals, can help estimate users' emotional states and readiness to engage, enabling systems to adapt the timing and style of arguments. The talk also addressed group scenarios, where emo-tional differences among users affect consensus formation and where agent design can support balanced discussions. Finally, ethical considerations were emphasized, partic-ularly when persuasive systems interact with vulnerable users such as children. These perspectives suggest a framework for emotion-aware, personalized, and trustworthy argumentative recommender systems.

Evaluation and Impact of Argument Based Recommendations

Mark Colley - University College London, UK

Argument based recommendations make the reasoning behind a recommendation explicit and open to inspection, contestation, and refinement. Rather than optimizing only acceptance, their impact should be assessed through decision quality, confidence calibration, transparency, user understanding, and appropriate reliance. In the talk, I discussed practical evaluation strategies, including controlled user studies, simulation based methods, and trade off analysis across multiple outcome measures (e.g., using Bayesian Optimization). I also highlighted that effects are often heterogeneous: some explanation styles support credibility and trust calibration, while weak or generic reasons can increase effort or even backfire. Finally, I connected these points to my own work on human AI interaction in automated vehicles and adaptive systems, where understanding user reasoning and behavior remains a central challenge.

Better Recommendations through Variety - News Recommendations and VAAs

Abraham Bernstein - University of Zurich, Switzerland

Abstract not available.

Ethical Dimensions for Recommender systems

David Traum - University of Southern California, USA

Abstract not available.

Architectural Perspectives on LLM/SLM-based Recommendation Systems - Edge, Distribution, and Privacy as Design Constraints

Keiichi Yasumoto - NAIST, Japan

Recent studies on LLM-based recommender systems have explored the generation of explanations and arguments, typically assuming centralized large-scale models. However, real-world systems operate under strict constraints such as latency, limited computational resources, and strong privacy requirements, making system architecture and deployment as important as model design. This talk addresses two fundamental questions: (Q1) what architectures and representations are suitable for integrating LLMs and argumentation in recommender systems, and (Q2) what types of data are required to support argument generation and evaluation under privacy constraints. Argumentation is framed not only as a modeling problem, but also as an architectural one.

A shift from centralized LLMs to distributed systems based on small language models (SLMs) is considered, where edge models capture user context and personal signals, local or fog nodes perform candidate generation, and central models are invoked for complex reasoning when necessary. Empirical examples, including IoT device control and on-device user context understanding, demonstrate that meaningful reasoning and human-readable explanations can be achieved locally without relying on centralized processing. As recommender systems increasingly aim to support user understanding, trust, and behavior change, a key challenge lies in determining where argument generation and evaluation should be performed, with privacy constraints playing a central role in shaping how and where argumentation is realized in next-generation recommender systems.

Evaluation & Impact of Argument-Based Recommendations

Vera Schmitt - TU Berlin / DFKI, Germany

Argument-based recommender systems (ABRs) promise to enhance transparency and user trust by grounding recommendations in structured evidence and natural language rationales. Yet evaluating their quality and downstream impact remains a multidimensional challenge. This lightning talk proposes a layered evaluation stack encompassing automatic retrieval metrics (Precision@k, nDCG, MRR), faithfulness tests via perturbation-based methods, human-centered argument quality dimensions (relevance, sufficiency, coherence), and controlled user studies measuring decision performance, trust calibration, reliance, and cognitive load. Drawing on empirical work in AI-supported fact-checking with lay users (N ¿ 400) and experts (N = 27), we show that argument-style natural language explanations significantly improve subjective dimensions (understandability, perceived usefulness, and trust), while failing to consistently improve objective decision accuracy and introducing increased variance in human performance. Critically, explanations can induce over-reliance on the AI system, with users deferring to argument-supported recommendations even when they are wrong. We further examine the role of LLM-as-a-Judge (LaaJ) as an evaluation proxy, presenting evidence that LaaJ and human evaluation capture largely orthogonal quality signals, with fewer than 8% of surveyed papers validating this equivalence directly. We conclude that no single metric suffices: the subjective appeal of argument-based explanations must be weighed against their objective impact on decision quality, necessitating stratified evaluation designs that combine automatic, LaaJ, and human methods.

Don't Dump - Guide: How AI Should Explain Itself Over Time

Elisabeth André - University of Augsburg, Germany

Abstract not available.

Concept of "Information Health" - Balancing Personalization and Diversity

Yutaka Arakawa - Kyushu University, Japan

At the Shonan Meeting, I presented the concept of "Information Health" — a state in which users' right to access diverse and balanced information is actively protected

against the distorting effects of algorithmic personalization. Using a food-and-nutrition analogy, I argued that today's recommendation systems risk producing "information obesity": just as a diet of only junk food harms the body, a steady feed of algorithmically filtered content narrows our worldview.

To address this, I introduced four research initiatives from my lab: a filter bubble detection system that measures browsing diversity and recommends topically adjacent articles; an AI-generated diverse comment system that exposes readers to a spectrum of viewpoints on news; a political stance-controlled agent framework for producing ideologically calibrated AI personas; and a Chrome extension that nudges users away from compulsive YouTube consumption through bias badges and visual de-emphasis.

My core argument is that AI must evolve from a waiter — serving more of what you already consume — into a registered dietitian that actively promotes a balanced and healthy information diet.

Hybrid Argumentation and Responsible AI

Khalid Al-Khatib - University of Groningen, The Netherlands

Large Language Models offer new opportunities for education, research, and decision support, but they also produce unsubstantiated claims, hallucinated facts, and misinformation, which prompts concerns about epistemic integrity. In this talk, I introduced a hybrid argumentation interface that enables structured, collaborative reasoning between humans and AI, and I framed it as a direction for recommendation tools that go beyond suggesting options to providing reasons, exposing trade-offs, and adapting to user constraints. The key question is how humans and AI can jointly engage in argumentative dialogue to improve reasoning and recommendations. The interface combines different argumentation theories to make both AI outputs and user responses explicit and contestable, supporting arguments, counterarguments, and targeted follow-up questions rather than generic advice.

Ethics, Manipulation & Responsible Deployment

John Lawrence - University of Dundee, UK

Abstract not available.

Evaluation & Impact of Argument-Based Recommendations

Martin Baumann - Ulm University, Germany

The presentation examines how to evaluate the quality and impact of argument-based recommender systems, with a particular focus on human cognitive processing. It highlights the limitations of traditional evaluation metrics and emphasizes the need to incorporate criteria related to how users perceive, understand, and process arguments. Drawing on theories of text comprehension and the Elaboration Likelihood Model (ELM), the work distinguishes between analytic (central) and heuristic (peripheral) processing routes. It proposes measuring both the depth of argument processing (e.g., recall, effort, sensitivity to argument strength) and the resulting attitudinal and behavioral outcomes. Empirical insights from user studies on conversational agents illustrate how communication style and disclaimers influence trust and persuasiveness.

Overall, the presentation argues that high-quality argument-based recommendations should foster deeper cognitive engagement and lead to more durable and meaningful user impact, which in turn can be utilized to measure the impact and quality of argument-based recommender systems and their recommendations.

Summary of discussions

As mentioned above, the discussions were organized around three central research themes: Group 1 examined the ethical dimensions of argument-based recommender systems, with particular attention to the boundary between nudging and manipulation, and proposed a value-driven architecture for responsible deployment. Group 2 considered the methodological challenges involved in evaluating such systems, arguing for a process-based approach that captures both recommendation quality and argument faithfulness across different levels of impact. Finally, group 3 explored questions of personalization and interaction design, emphasizing the complexity of multi-user settings and deliberative interactions, especially in sensitive domains such as healthcare.

The following sections summarize the main outcomes of each group's discussion.

Group 1: Ethics, Manipulation & Responsible Deployment

Explanations clarify why something occurred or is the way it is and aim to increase understanding. Arguments, however, look at why a claim (or a recommendation) should be believed or endorsed and aim to increase conviction or knowledge. Thus, argument-based recommender system using their argument map in an interactive manner with the user are ethically more concerning.

The discussion of this group was guided by the following questions:

1. When does argumentative recommendation cross into manipulation or undue persuasion?
2. How can we design argument-based systems that promote autonomy, fairness, and informed decision-making?
3. What are the risks and benefits of deploying argumentative recommender systems in sensitive domains like health, education, or politics?

The overarching conclusion of the discussion is that we need to look at the content of interactive, argumentative recommender systems from a value-oriented perspective. With an explicit value system, we can analyze argument-based recommender systems concerning manipulation and therewith create justified trust. The value system that is in place decides whether actions are ethically justifiable.

The value-driven argument-based recommender system that the groups has envisioned is depicted in Fig. 2 and includes the following modules:

- User Model with task-related data, self-awareness, emotional regulation / affect information, value system
- Recommender System with value system
- Domain Argument Map that is providing arguments for and against the recommendation
- Meta-Reasoning Argument Map for the usage of the domain argument map

Given this setup, the group gathered the following ethical conundrums:

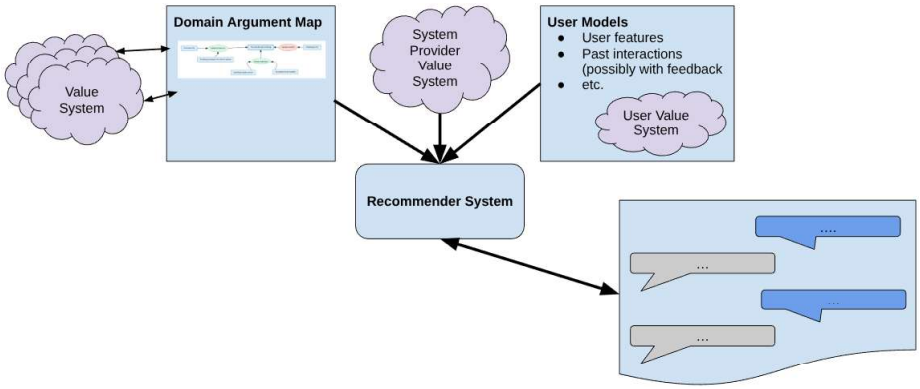


Figure 2: Value-driven Argument-based Recommender System

- *Nudging* is emphasizing parts of the map towards a specific goal and therewith subtlety influences people's decision-making, whilst still allowing them to choose freely. In the context of argument-based recommender systems this means systematically choosing a specific path through the argument map when making supporting or refuting arguments for a recommendation.
- *Manipulation/Steering* is systematically and/or internationally hiding/ withholding parts of the argument map towards a specific goal. In contrast to law, we call it manipulation/steering independent of whether the pursued goal is "good" or "bad". In the context of argument-based recommender systems this means systematically choosing or omitting arguments from the map when supporting or refuting a recommendation.
- *Value Conflicts/Tradeoffs* can be made explicit by weighing the different value systems against each other.

Additionally, the group gathered the following design principles:

- *Know what you don't know*: The argument map might help the system to recognize its limitations and inform users appropriately.
- *Limit Harm - Responsibility*: The meta-reasoning argument map may help reason about harmful recommendation.
- *Make the value systems explicit*: Allows for reasoning about tradeoffs and allows users to change/override their value system.
- *Make the value system dynamic*: We all change and might use the system over time and not only once (social relations, location, context, ...). Check for consistency and make the updates explicit.

Limitations and challenges of their approach include:

- dependence on the quality of the argument map and user model
- analysis does not ensure ethical compliance
- tacit information is difficult to make explicit

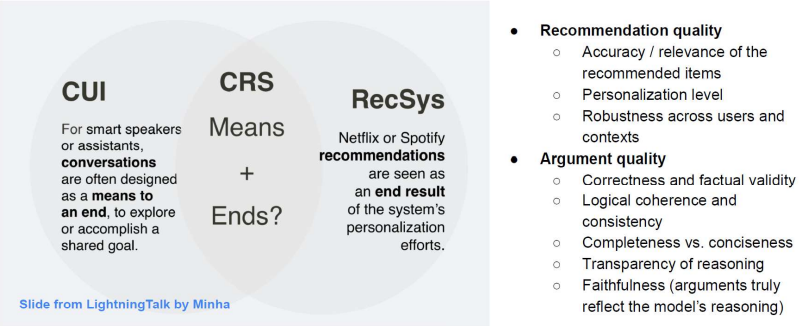


Figure 3: Main levels of impact considered in the evaluation of argument-based recommender systems.

- maintaining explicit information is costly and implicit information is cheaper to gather
- may not be in the interest of the stakeholder to make everything explicit (e. g., trade secrets)

However, their approach is generalizable, allows to interrogate recommendations, and the explicitness of the value systems allows for transparency.

Group 2: Evaluation of Impact of Argument-Based Recommendations

Evaluation of argument-based recommender systems requires a context-sensitive and risk-aware approach. The criteria used to assess system quality depend on the domain, the decision space, and the possible consequences of error. In high-risk settings, such as healthcare, evaluation must prioritize correctness, safety, robustness, and verifiability. In lower-risk or more open-ended settings, qualities such as usefulness, engagement, novelty, or support for reflection may become more important. Evaluation therefore cannot be reduced to a single benchmark or a universal metric, but must instead be aligned with the specific goals and risks of the application. In the Fig. 3 the levels of impact for conversational recommender systems is depicted.

Evaluation Dimensions. A broad evaluation perspective is needed because these systems do more than recommend items: they shape how users understand, compare, and justify choices. Assessment must therefore cover both recommendation quality and argument quality. Recommendation quality includes aspects such as relevance, personalization, and robustness across users and contexts (see Fig. 4). Argument quality includes factual validity, logical coherence, completeness, conciseness, transparency, and faithfulness to the actual reasoning process (see Fig. 5). In addition, evaluation should consider several levels of impact, including the technical properties of the system, the quality of the communication it produces, its effects on the individual user, and its wider societal consequences.

Process-Based Evaluation. Outcome alone is not sufficient as an evaluation target. It is not enough to know whether a recommendation was accepted or whether the final

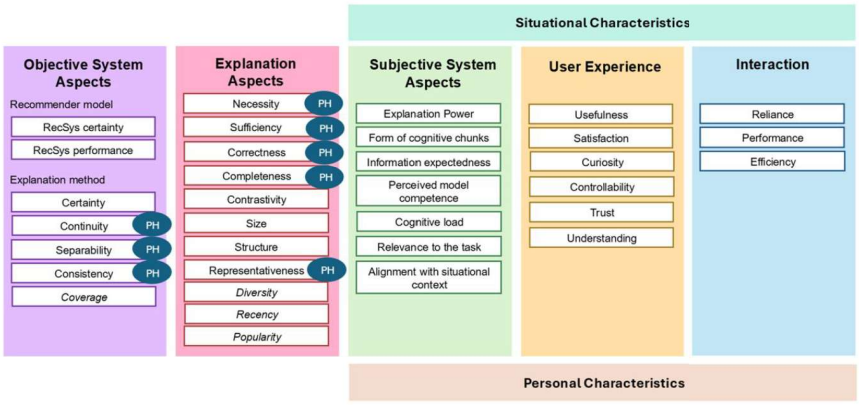


Figure 4: Example of a multidimensional quality assessment framework, covering system aspects, explanation aspects, user experience, and interaction. Image by Wardatzky et al. 2025 (https://ceur-ws.org/Vol-4063/paper2.pdf).

answer appears correct. What matters equally is whether users engage with the arguments, compare alternatives, understand the reasons behind the recommendation, and refine their view of the decision space. For dialogue-based systems, this makes multi-turn and process-based evaluation especially important. The quality of the interaction, the degree of user engagement, and the effect on the user's mental model are all central parts of impact assessment.

Trust, Manipulation, and Safety. A major concern is the distinction between trust and trustworthiness. A system may sound confident and persuasive without actually deserving reliance. Self-reported trust is therefore an incomplete indicator of quality. Evaluation must also address manipulation risks, including biased or toxic data, adversarial influence, and system behavior that subtly steers users while appearing neutral. These issues become more important when the system presents arguments in a convincing style, since persuasive fluency can mask weak reasoning or unreliable evidence. Safety assessment must therefore include robustness against misuse as well as scrutiny of how arguments are generated and presented.

Long-Term and Societal Impact. Impact assessment must extend beyond immediate interaction. Repeated use of argument-based recommender systems may affect users' dependence on automated advice, their willingness to think critically, and their reliance on machine-generated justifications instead of human discussion. At a broader level, widespread adoption may influence social norms, reduce diversity of viewpoints, or encourage convergence toward similar forms of reasoning and decision-making. These longer-term and collective effects are difficult to observe through short studies, but they are essential for responsible evaluation.

Methodological Implications. An appropriate methodology must combine several perspectives. It should be adaptable to different use cases, sensitive to stakeholder needs, and capable of examining both ordinary and edge cases. Expert evaluation may

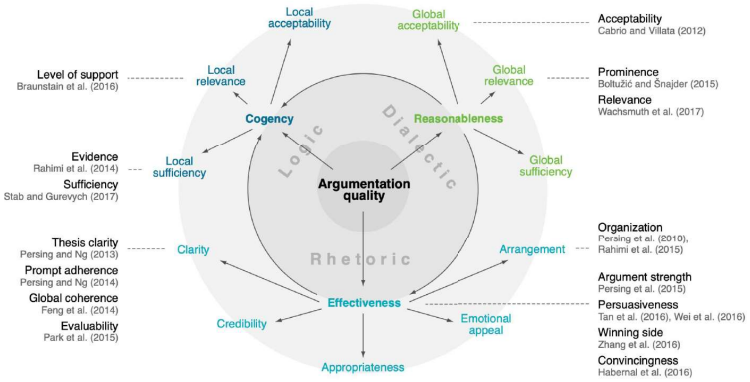


Figure 5: Illustration of argument quality dimensions, including clarity, relevance, sufficiency, credibility, effectiveness, and acceptability. Image by Wachsmuth et al. 2017 (https://aclanthology.org/E17-1017.pdf).

be necessary in high-stakes domains, while user studies remain important for understanding interaction quality, perceived usefulness, and behavioral effects. Adversarial testing, value-oriented analysis, and long-term impact studies are also needed. Taken together, these considerations position evaluation as a central research challenge for argument-based recommender systems, requiring both technical analysis and broader socio-ethical assessment.

Group 3: Personalisation and Interaction Design

Personalisation as the Basis of Argumentative Recommendation. A central conclusion of the discussions was that effective argumentative recommender systems depend on rich personalization. The quality of recommendations and explanations improves when the system can take into account a user's background, preferences, values, interaction history, and immediate context. This applies not only to what is recommended, but also to how arguments are framed. Different users may respond to arguments based on health, environmental responsibility, personal benefit, or social harmony, and these preferences are often shaped by both individual traits and cultural context. At the same time, this degree of personalization raises serious privacy concerns, especially when systems rely on sensitive personal or health-related information.

Interaction Design and Proactivity. Another major theme concerned how systems should manage interaction. Proactive behavior was understood as offering information, explanations, or warnings without being explicitly asked, and the discussions made clear that this should be carefully calibrated to the situation. In safety-critical contexts, proactive intervention may be necessary, while in more subjective situations the system should remain open to negotiation and dialogue rather than acting too assertively. Excessive proactivity can easily become intrusive, tiring, or inappropriate, so interaction design should allow systems to adjust to the user's needs, the state of the conversation, and the seriousness of the context. At a minimum, users should be able to limit or disable proactive behavior.

Multi-User Settings and Consistency. The discussions also highlighted the complexity of recommendation in group or multi-user settings. A system may need to adapt its language and arguments to different users, but if it changes too much across interactions, it risks appearing inconsistent or untrustworthy. This creates a tension between adaptation and stability: systems must remain flexible enough to address different people while also maintaining a coherent identity. In such contexts, argumentation also depends on source credibility, cultural sensitivity, and awareness of shared or conflicting goals. Supporting multi-party recommendation therefore requires not only user modelling, but also an understanding of social dynamics and the relationships between stakeholders.

Healthcare as a Use Case for Personalized Argumentation. These ideas were explored through a healthcare use case focused on treatment deliberation. This setting was considered especially relevant because it involves multiple stakeholders, meaningful alternatives, and high demands for explanation and trust. Personalization in this context would need to include factors such as health condition, age, emotional state, knowledge level, decision-making style, and preferences for detail or privacy. The discussions also showed that the role of the system could vary considerably: it might act as an expert-like advisor, a supportive counselor, or a peer-like source of experiential knowledge. Beyond recommending options, such a system could also help users prepare arguments or questions for doctors, family members, or insurers.

Multiple Perspectives and Deliberative Interaction. A further insight was that conflicting viewpoints may be easier to understand when they are presented through distinct identities rather than merged into a single neutral summary. Multi-agent or multi-persona systems can make disagreement more visible, encourage critical reflection, and reduce the burden on the user by allowing agents to challenge one another directly. This design supports deliberation by exposing alternatives and counterarguments in a more engaging and interpretable way. However, the discussions also noted important challenges, including information overload, the risk of weak context sensitivity, and the difficulty of matching argumentative strategies to different users.

Trust, Evaluation, and Design Implications. Finally, trust emerged as a key design concern. Trust was treated as multifaceted and context-dependent, shaped not only by system behavior but also by source transparency, institutional affiliation, prior beliefs about technology, and even the visual or vocal presentation of the agent. The discussions suggested that recommender systems should not simply aim to be trusted automatically, but should instead support informed and critical engagement. Overall, the main design implication is that personalized argumentative systems should be sensitive to users, context, and culture, while remaining transparent, restrained in their proactivity, and structured in ways that help users reflect rather than merely comply.

Key Takeaways and Research Directions. The discussion produced a broad framework for persona-driven argumentative recommender systems in healthcare. Important design elements include detailed modeling of patient characteristics, such as demographics, cognitive ability, affective state, cultural background, need for cognition, and familiarity with technology. In addition, systems must account for stakeholder relations, including who makes decisions, who is affected by them, who pays, and who influences the outcome. The role adopted by the system may also differ depending on

the situation, for example acting as a knowledge-focused advisor, an experience-based peer, or an emotionally supportive counselor. Possible outputs extend beyond treatment recommendations to include verification, analysis, and questions that patients can ask in later consultations. These functions must be grounded in careful attention to privacy, liability, and the kinds of data available, such as diagnosis, clinical records, and patient history.

Several open research questions remain. A major challenge is how to match argumentation strategies to individual user characteristics, including expertise, decision-making style, and preference for detailed reasoning. Other unresolved issues include how to manage conflicting stakeholder interests, how to measure trust in ways that go beyond simple questionnaires, and how to maintain user engagement during long deliberations without losing depth. Further questions concern the validation of synthetic personas for evaluation, the handling of cross-cultural medical recommendations in settings where legal and insurance systems differ, and the need for mechanisms that allow users to question or challenge system-generated conclusions.

From an implementation perspective, the discussion suggested that voice-based interaction may make deliberative systems more engaging than text alone, while moderator components are essential for preventing premature agreement and encouraging critical thinking. Good system performance also depends heavily on context awareness, including location, lifestyle, and personal circumstances. At the same time, legal and regulatory requirements vary across countries and must be reflected in the design. Techniques that deliberately encourage deeper reflection may improve decision quality, even if they reduce immediate user satisfaction, and interaction styles that allow users to gradually "join in" may reduce pressure in sensitive domains such as healthcare. A multi-agent deliberation platform was presented as a promising research environment for exploring these issues, since it can support multiple personas and strategies, moderator-driven interventions, structured summaries of agreement and open questions, and future extensions to voice-based interaction.

Summary of new findings

1. Importance of Value-Driven and Context-Aware Personalization The group discussions showed that effective argument-based recommender systems should be grounded in explicit value systems and rich forms of user personalization. Group 1 emphasized that making value systems explicit enables users to examine trade-offs and develop justified trust in the system. Group 3 complemented this perspective by showing that recommendations and explanations become more effective when they are adapted to a user's background, emotional state, and immediate context. Taken together, these findings suggest a shift away from generic system outputs toward highly tailored and value-sensitive argumentative interactions.

2. Shift Toward Process-Based and Multi-Dimensional Evaluation The discussions also made clear that outcome-based metrics alone are insufficient for assessing argumentative systems. Group 2 argued that evaluation must account for the full reasoning process, including user engagement, the quality of multi-turn interaction, and the system's effect on the user's mental model. In addition, evaluation should adopt a multi-dimensional perspective that distinguishes between the quality of the recommendation itself, such as relevance and robustness, and the quality of the arguments, such as logical coherence and factual validity.

3. The Value of Multi-Persona and Deliberative Interaction Another important finding concerned the benefits of presenting conflicting viewpoints through distinct synthetic identities rather than through a single neutral summary. Group 3 argued that multi-agent platforms can make disagreement more visible, understandable, and manageable for users. This design supports deliberation by dynamically exposing alternatives and counterarguments, which appears particularly valuable in multi-stakeholder settings such as healthcare.

Across all groups, there was broad agreement that current systems still struggle to distinguish superficial persuasiveness from genuine trustworthiness. The discussions highlighted the difficulty of separating beneficial nudging from unethical manipulation, especially when persuasive fluency conceals weak underlying reasoning. At the same time, all groups pointed to a persistent tension between the need for extensive user data to support effective personalization and the resulting privacy risks, liability concerns, and costs of maintaining explicit user models.

Identified issues and future directions

Identified Issues

The discussions at the Shonan Meeting highlighted several important issues. One central problem concerns the boundary between manipulation and persuasion: systems that appear confident and helpful may cross from legitimate nudging into subtle manipulation when they systematically hide, omit, or emphasize selected parts of an argument map. A second issue is the limitation of current evaluation approaches, which often fail to capture long-term societal effects such as growing dependence on automated advice or the narrowing of human perspectives and forms of reasoning. A further challenge lies in the dependence of high-quality personalization on sensitive information, including emotional, behavioral, or health-related data, which creates serious privacy concerns and makes the explicit maintenance of such information costly. Finally, in multi-user settings, systems must balance adaptation to individual stakeholders with the need to maintain a coherent and trustworthy identity, while also avoiding excessive complexity or information overload.

Future Directions

The discussions also pointed to several promising directions for future research. One important direction is the development of dynamic and user-accessible value systems, together with meta-reasoning argument maps that can reflect changes in user context, relationships, and priorities over time. Another priority is methodological innovation in evaluation, including hybrid approaches that combine expert assessment in high-stakes domains, adversarial testing for safety and robustness, and longitudinal studies of broader societal impact. A further open challenge is the design of systems that can adapt argumentative strategies to individual user characteristics, such as decision-making style, expertise, or need for cognition. Finally, the development of multi-agent deliberation environments, potentially supported by moderator components and voice-based interaction, appears to be a promising path for encouraging critical reflection, preventing premature agreement, and creating more natural forms of argumentative dialogue.

Introductory slides of Participants

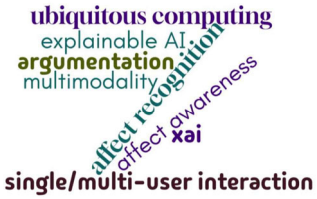
Matthias Kraus - LMU, Germany

Yuki Matsuda - Okayama University, Japan

Wolfgang Minker - Ulm University, Germany

Wolfgang Minker – Dialogue Systems Group Ulm University (Germany)

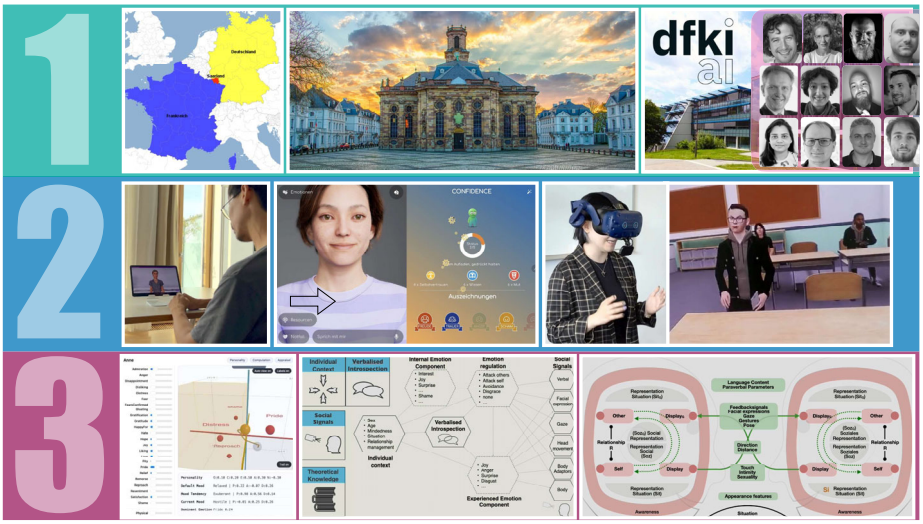
- Proactive user- and situation-adaptive dialogue systems
  - Cooperative multi-user interaction
  - Modelling of argumentative dialogue and learning appropriate strategies
  - Culturally adaptive dialogue management
- Explainable AI
  - Interactive explanations of machine learning methods through dialogue systems
- Affective Computing
  - Automatic Recognition of Psychophysiological States Based on Multimodal Data Analysis
  - Exploration of Affective Computing in Adaptive Multimodal Dialogue Systems



- Expectations
  - New perspectives on recommender systems
  - Clearer sense of open research gaps
  - Potential collaborators

dialogue systems universität ulm

Patrick Gebhard - DFKI Saarbrücken, Germany



Carolin Schindler - Ulm University, Germany

M.Sc. Carolin Schindler

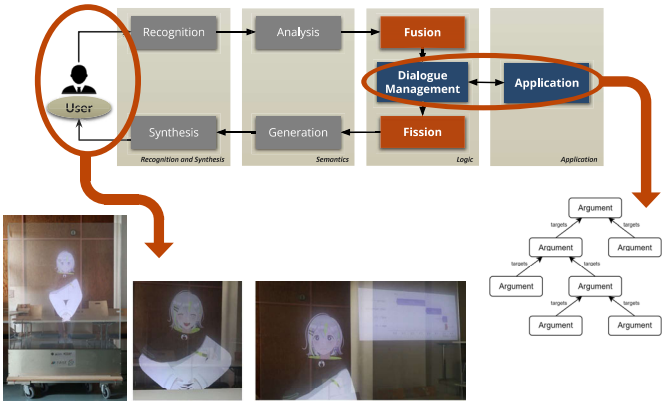
Academic Background

- Bachelor of Science: Communications and Computer Engineering
- Master of Science: Cognitive Systems

Double-Degree PhD Student (Ulm University & NAIST)

Natural Multimodal Interaction with Argumentative Dialogue Systems

carolin.schindler@uni-ulm.de
https://nt.uni-ulm.de/schindler



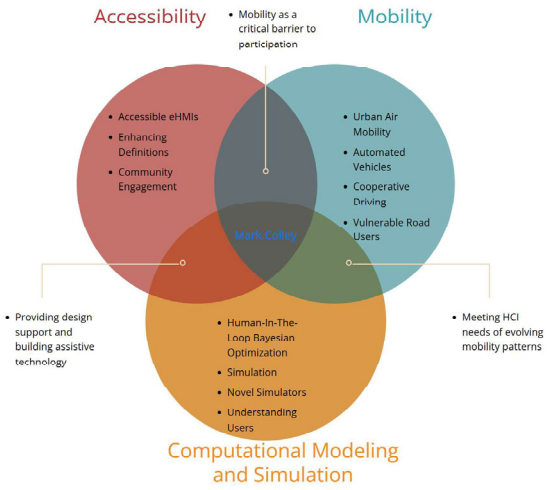
NAIST

universität ulm

Mark Colley - University College London, UK

UCL

東京大学 THE UNIVERSITY OF TOKYO

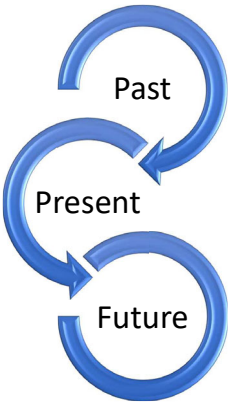


Mark Colley

Lecturer / Assistant Prof in HCI

https://m-colley.github.io/

Jacqueline Urakami - Kyocera, Japan



Academia (20+ years)
Cognitive and Engineering Psychology
Human-Centered Design
Human-Machine Interaction Design

Kyocera R&D (4+ years)
Human Augmentation
Research on Vibrotactile Feedback for Memory and Emotion
Team Leader "Educational Technology / Haptics"

Future Directions
Human-(in context) Centered Design
Empathy as cognitive skill
Working with AI, not designing for AI
Ethics as a Design skill (human control, privacy, trust)

Elisabeth André - University of Augsburg, Germany

Elisabeth André

Founding Chair of Human-Centered AI @University of Augsburg, Germany

- **My Interests**:

Recommender Systems

CARE
Context-Aware Recommender System for the Elderly
2014 - 2016

IT4SE
IT for Smart renewable Energy generation and use
2010 - 2015

Trust

TRUST
Trustworthiness of Organic Computing Systems
2009- 2015

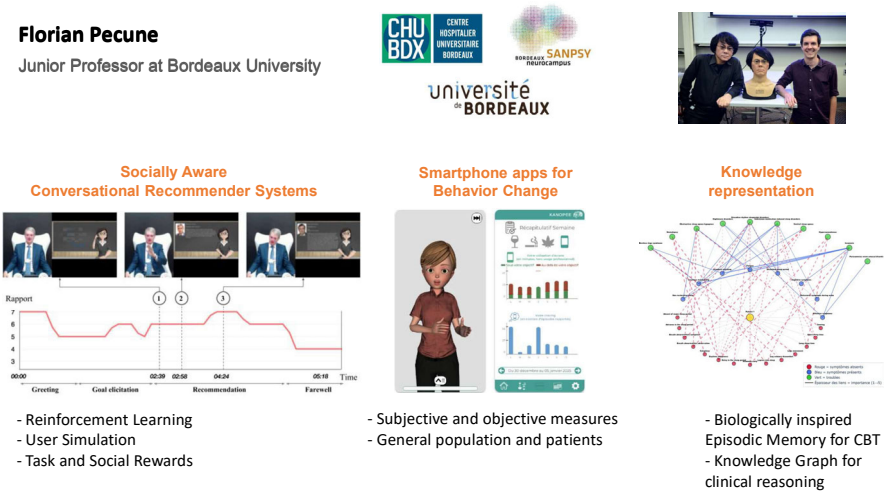
Argumentation

BEA
Building Engaging Argumentation
2021 - 2024

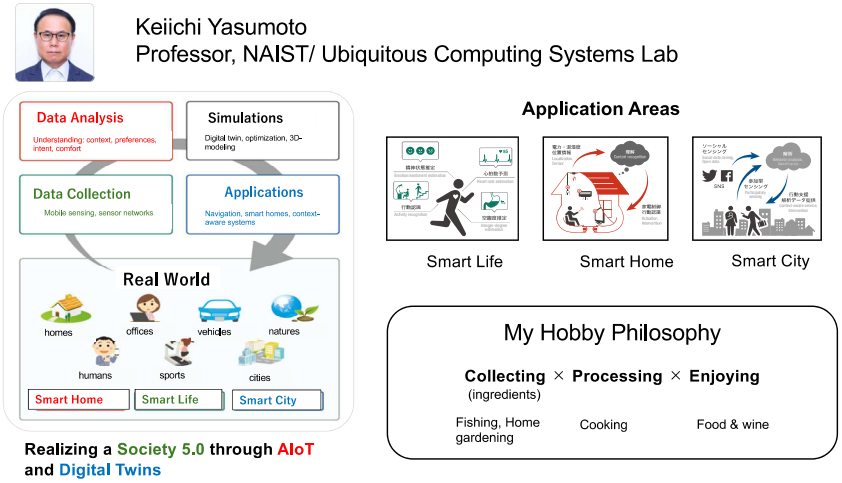
EVA
How to Win Arguments - Empowering Virtual Agents to Improve their Persuasiveness
2018 - 2021

- **What I expect:**
  - **New connections**
    Between trust, argumentation, and recommender systems - and between the people working on them.

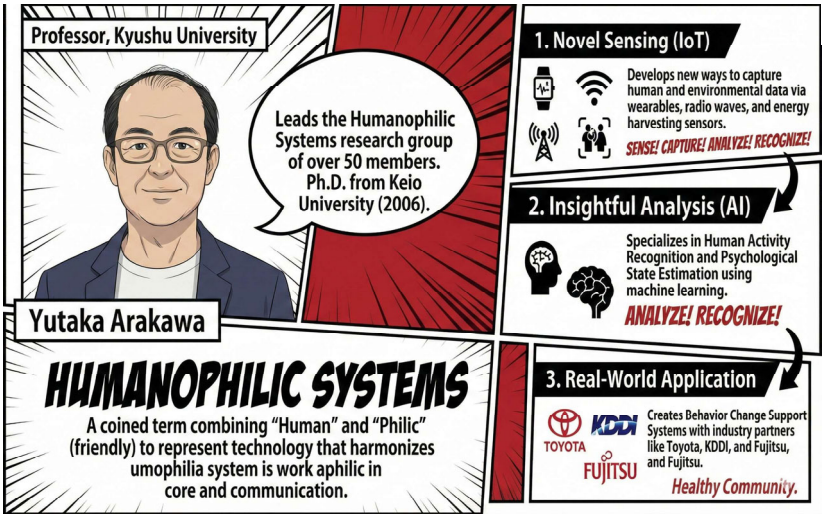
Florian Pecune - University of Bordeaux, France



Keiichi Yasumoto - NAIST, Japan



Yutaka Arakawa - Kyushu University, Japan



Kaoru Sumi - Hakodate Future University, Japan

Kaoru Sumi

Affiliation / Position:

- **Professor, Future University Hakodate**
- General Chair, Persuasive Technology 2026
- Guest Editor, Frontiers in AI – "Next-Generation Persuasive Technologies"
- Editorial Board: JoVE; Interactive Technologies for Behavior Change and Emotional Engagement
- Associate Editor: Behaviour & Information Technology

From / Background:

Originally from Tokyo, Japan

Research Interests:

- Human–Agent Interaction
- Persuasive Technology & Behavior Change
- Affective Computing
- XR / MR / Metaverse Interaction

Topics of Interest:

- Personalized persuasive experiences
- XR agents & embodied interaction
- Emotion-driven adaptive interfaces
- Data-driven behavior modeling

Recent Representative Research:

- Nonverbal information–based emotion inference and adaptive interaction design
- XR-based interaction design and learning support systems
- Phantom sense & embodiment experiments with EEG/ECG
- Emotion expression and affective motion design for four-legged virtual pets

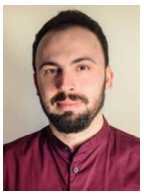
Contact:

Email: kaoru.sumi@acm.org

Website: https://www.fun.ac.jp/en/faculty/sumi-kaoru

PAHAI Lab | Persuasive & Affective Human–AI Interaction

Khalid Al-Khatib - University of Groningen, The Netherlands



Khalid Al-Khatib

Assistant Professor
University of Groningen, The Netherlands

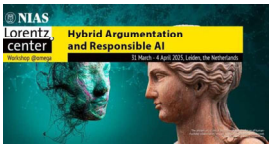
https://www.rug.nl/staff/khalid.alkhatib
https://dblp.org/pid/31/8936.html

My research Areas

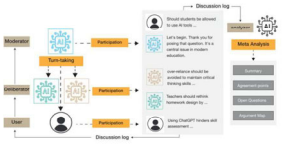
- Argument mining and assessment
- Argumentation knowledge graphs
- Persuasion & deliberation strategies
- Hybrid Argumentation and Bias analysis

Why excited about the Seminar

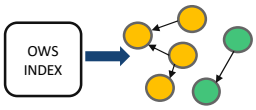
*It's not only about better recommenders; it's about helping **people think better**: ask why, challenge the system, and make more informed choices. A win-win :-)*



Hybrid Argumentation

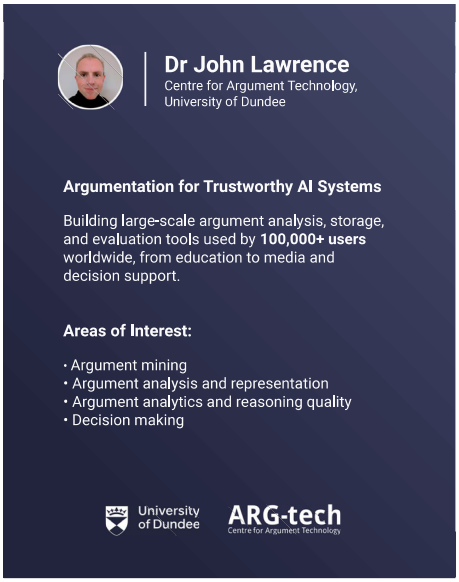


ArgsBase



AKASE

John Lawrence - University of Dundee, UK

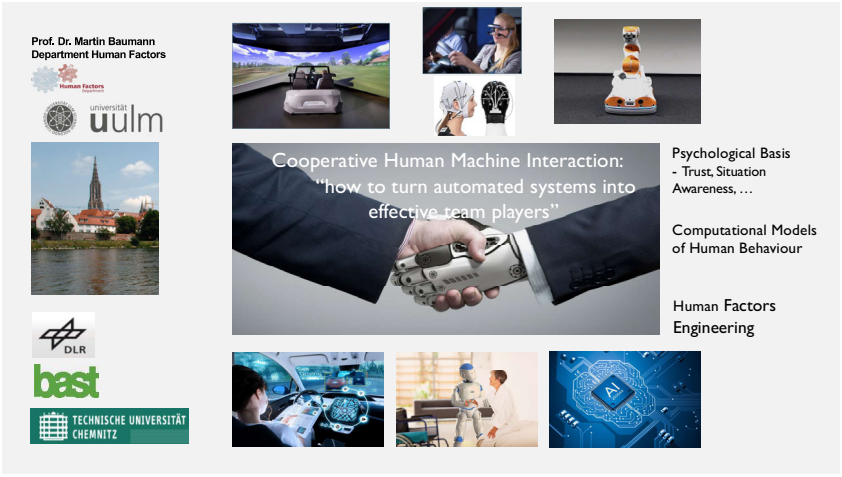


BBC
EVIDENCE TOOLKIT

AIFdb

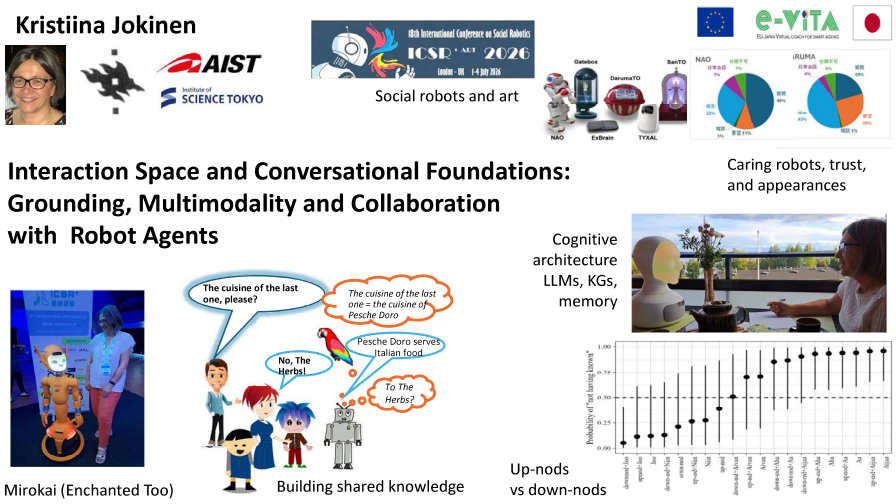
open
Argument
Mining
Framework

Martin Baumann - Ulm University, Germany



Vera Schmitt - TU Berlin / DFKI, Germany

Kristiina Jokinen - AIST, AIRC, Japan



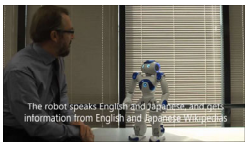
Graham Wilcock - University of Helsinki, Finland

Graham Wilcock

- Social Robotics
- Conversational AI
- Knowledge Graphs

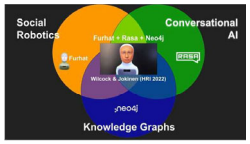
graham.wilcock@helsinki.fi
Adjunct Professor, University of Helsinki
Visiting Professor, Doshisha University 2015-16
Visiting Professor, Kyoto University 2018-19


Finnish WikiTalk, Helsinki 2015

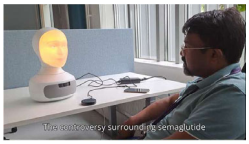


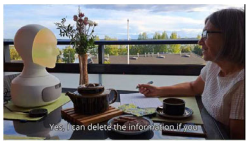
Japanese WikiTalk, COLING 2016

ERICA and WikiTalk, IJCAI 2019

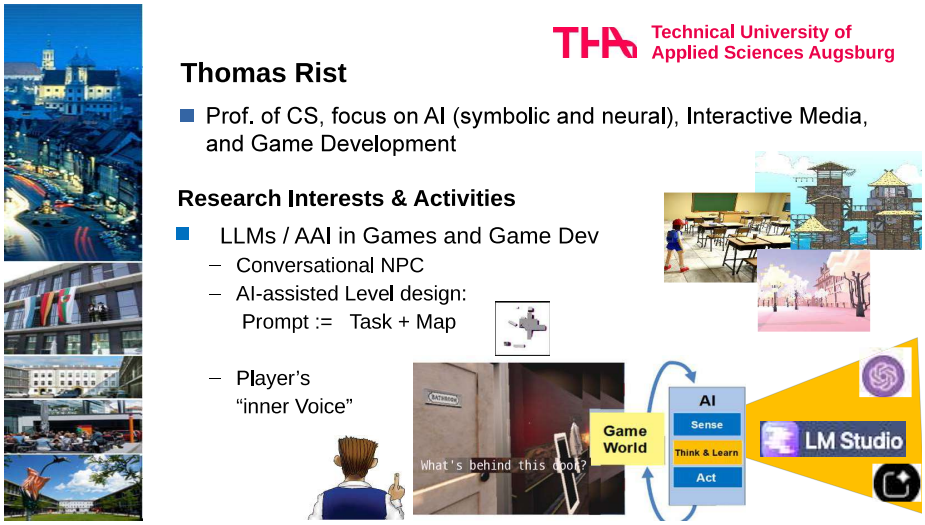


CityTalk, HRI 2022, RO-MAN 2023

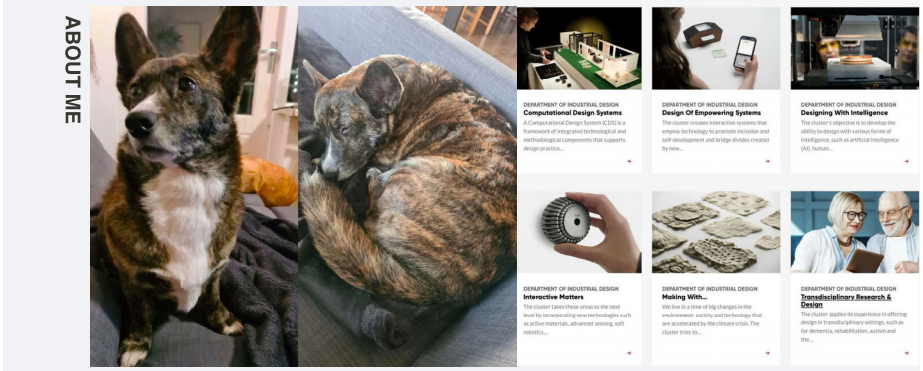

Expert Interaction, ICSR 2025


Episodic Memory, ASIMOV 2025

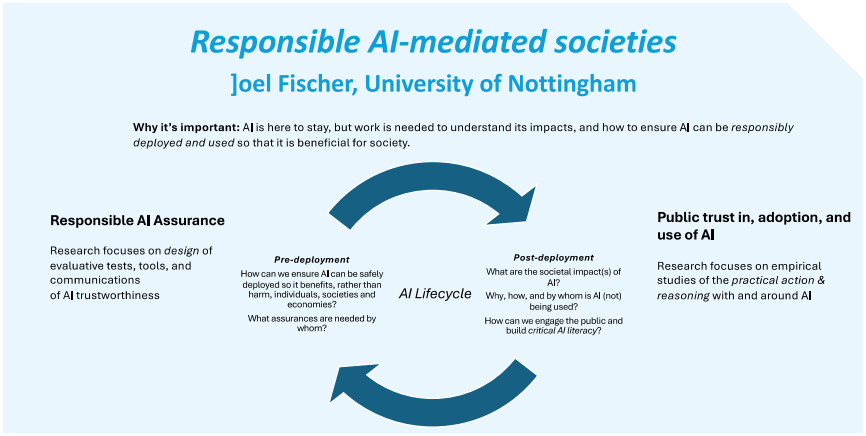
Thomas Rist - Augsburg University of Applied Sciences, Germany



Minha Lee - TU Eindhoven, The Netherlands



Joel Fischer - University of Nottingham, UK



Patricia Kahr - University of Zurich, Switzerland

PATRICIA KAHR

University of Zurich — **Postdoctoral Researcher**
Dynamic & Distributed Information Systems Group

IMPERIAL — **Visiting Researcher**
I-X Inititiative / Human-Robot Interaction

TU/e — **PhD Candidate**
Human-Technology Interaction Group; Thesis: *The Dynamics of Trust and Reliance in Human-AI Interactions*

Maastricht University — **MSc Human Decision Science**
Thesis: *Gender Stereotypes in Human-Algorithm Interaction*

- (dyadic, group) ***decision-making*** scenarios with AI
- focus on ***user perspective:*** understanding ***perceptions*** (trust, comprehension , decision control, meaningfulness, self-efficacy) and ***behavior*** (appropriate reliance / deferral, decision performance, learning)
- ideally: studying HAI develop in long-term ***setups***
- ***people + AI,*** *but also:* ***people > AI***

- AI support in **multi-stakeholder** / **group** scenarios
- **Proactive** AI in HAI for better human-AI alignment*
- Discussions on TRUST and its role in our research field

* minus the persuasion aspect

my cat ***Elmo***, ***exploring cities on foot & countries by car***, *crafts*, movies, good iced coffee – with good people

Ko Watanabe - RPTU Kaiserslautern/DFKI, Germany

NII Shonan Meeting - Building Trustworthy and Interactive Recommender Systems through Argumentation

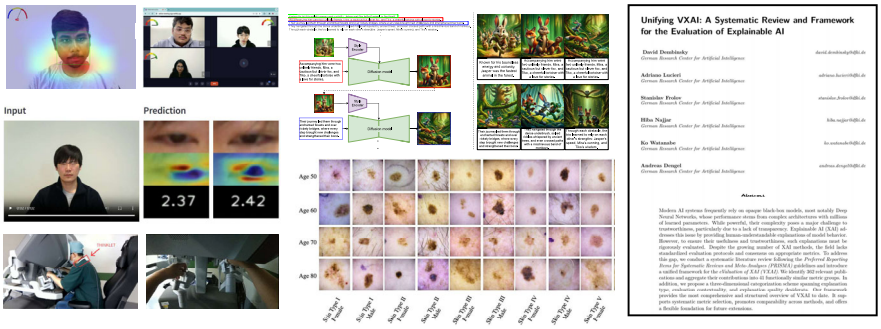
Dr. Ko Watanabe

dfki Deutsches Forschungszentrum für Künstliche Intelligenz German Research Center for Artificial Intelligence

OMU Osaka Metropolitan University

DFKI GmbH (KL, Germany), Visiting Researcher OMU (Japan)

Research Interest - Activity Recognition, Cognitive Augmentation, Medical/Healthcare AI



Abraham Bernstein - University of Zurich, Switzerland

Abraham Bernstein, University of Zurich

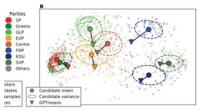
Background

Computer Science and Management

- **Education**
  - ETH zürich
  - Massachusetts Institute of Technology
- **Employment**
  - NYU
  - Universität Zürich
- **Sabbaticals**: CMU, U Syndey, U of Queensland

Goal

Solve societally relevant scientific problems



Research

Bridge AI/CS with social science and regulation

- Combining human and machine intelligence
- CSCW / CI
- Digital Direct Democracy
- Knowledge Graphs
- Neuro-symbolic AI
- Recommender Systems
- Trust/reliance in AI

Exchange

What I would like to talk about (excerpt)

- Explanations in RecSys
- Deliberation (in democracy & RecSys)
- Anything human-related RecSys
- …

and anything thought-provoking

Adaptive political surveys and GPT-4: Tackling the cold start problem with simulated user interactions, Bachmann F, van der Weijden D, Heitz L, Sarasua C, Bernstein A (2025) Adaptive political surveys and GPT-4: Tackling the cold start problem with simulated user interactions. PLOS ONE 20(5): e0322690. https://doi.org/10.1371/journal.pone.0322690

David Traum. *traum@ict.usc.edu*
Director for Natural Language Research, USC ICT
Research Professor, USC Viterbi School of Engineering
Editor in Chief, Dialogue and Discourse
https://people.ict.usc.edu/~traum/

1. Relevant Prior Work Topics

- **Speech acts approach to Grounding:** how agents reach "Common Ground"
- **Negotiation & Persuasion:** Moving beyond simple task-completion to agents that can engage in **multi-party negotiation** and **argumentation** to resolve **conflicting** goals.

2. Virtual Human Recommender Projects

SASO-ST (2000s): A platform for tactical negotiation where users must convince virtual characters to change their minds

SimSensei (2010s): An agent designed for high-stakes healthcare interviews, building rapport and recognizing distress

Human-Swarm Fire Rescue (2020s): human operator & robot team tries to locate, warn and rescue people from fire

3. Current Interests for this Workshop

- **Role of Context**: change meaning and appropriateness of communications based on Differences in
  - Genre (e.g., chat, task performance, advice, negotiation),
  - Activity (news interview vs legal interrogation),
  - Domain (restaurant, product, route/directions, medical intervention, ...)
  - Action (recommendation vs, warning, request, threat, bribe)
  - Relationship (subordinate, peer, adversary,..)
- **Evaluation:**
  - how good is a recommendation in general,
  - how appropriate is it for the specific person and situation
  - how well is it presented
- **Role of Interaction**
  - Joint exploration/decision, rather than separate recommender and decision maker
  - Exposure to different points of view. Distinguish undisputed fact from opinion