Unit 3 - Seminar Preparation:

I have selected two research papers in the field of Artificial Intelligence (AI), each employing a distinct research methodology, which can be found in the list of references below. In the following I will evaluate whether and how both of them manage to achieve their research objective and stress strengths and weaknesses of their approaches in a structured way defined in the task description.

1. Familiarise yourself with the purpose, problem, objective, or research question of each paper. Are they in line with your experience or thoughts on the topic, contributing to the collective body of knowledge in this area?

Wang et al. (2023): This paper aims at gaining an understanding of how the bias that is present in AI-based decisions and the provision of explanations for the AI's suggestions affect the perception of fairness considering the decision. This provides an interesting enhancement to the collective body of knowledge as there currently is not a lot of literature regarding an evaluation of fairness of AI decisions. However, as biases in training data lead to biases in the models' results, answering this question is important, especially as AI gets more and more integrated in our daily (work) lives, e.g., by using GenAI for brainstorming.

Lai et al. (2023): This paper focuses on the empirical study design choices in human-AI decision-making research. It summarises research strategies in over 100 studies concerning decision tasks, AI assistance elements, and evaluation metrics. The objective is to structure and guide future research efforts in understanding and improving human-AI collaborative decision-making. This aligns with current interests in optimising the synergy between humans and AI systems, especially in critical

areas like healthcare and criminal justice. This contributes to the collective body of knowledge in AI and is especially important because we as society are currently figuring out how humans and AI best interact together, especially with regards to the usage of Generative AI and AI Agents.

2. Is the research methodology utilised in each paper appropriate for the stated purpose or question?

Wang et al. (2023): They conducted an experiment. Participants were asked to take on the role as travel agents and bid on housing for one night stays of their customers. They were provided with information on the real estates' features as well as with budget information about their clients. The client paid the bid price that the host had in mind and 50% of the difference between their budget and this bid price as reward for the travel agent. Thus, travel agents needed to make a good prediction on the host's bid price to maximise their own income. In some sub-groups within the experiment, AI predictions of the bid price together with explanations of how it derived these results were provided to the travel agents, differing in their bias levels. This study is generally well-suited for answering the given research questions, but also has some draw-backs. The advantage of the approach is that in a controlled environment it enables to assess the effect that different levels of bias of AI systems have on the perceived level of fairness of a decision made based on the system, providing interesting insights. However, while within this test scenario, different mechanisms can be analysed, humans tend to behave differently in a laboratory environment like in this study than they behave in their daily lives. Thus, the study's generalisation to general concepts could be limited.

Lai et al. (2023): This paper utilises a survey methodology, systematically reviewing empirical human-subject studies on human-AI decision-making. By summarising study design choices across multiple papers, the authors aim to provide a structured understanding of the field. This methodology is suitable for identifying current trends, gaps, and making recommendations for future research. Trends are defined by what topics are covered in the research and especially by evaluating which strategies are used in different research papers. Gaps are found by looking at questions that are still unanswered, even after going through an extensive amount of literature in this field, covering all proceedings of AI-related conferences on the topic of AI-human-interactions. However, it could gain value by also including papers from other sources, not only proceedings from conferences as these might provide an additional perspective. Recommendations for further research then directly follows from the open gaps that need to be filled by additional research.

3. In terms of data collection and analysis, is this also appropriate for the stated purpose or question?

Wang et al. (2023): Data is collected by watching human beings interact with each other in a close-to-real-world-scenario. While the drawback of a lack of generalisation prevails due to the fact that humans tend to behave different when watched, especially when in a laboratory environment. Thus, the data collection strategy could also be (in theory) improved by analysing real-world human interactions. However, even if not conducted in a laboratory environment, humans would still be expected to change their behaviour due to being observed.

Lai et al. (2023): The authors collect data from over 100 empirical studies and analyse them to summarise study design choices. The analysis involves categorising and evaluating aspects such as decision tasks, AI models, and evaluation metrics. This systematic approach is appropriate for understanding the landscape of human-AI decision-making studies and identifying areas needing further exploration. As explained above, however, the decision to only include papers from AI-related conferences could be reevaluated as a broader range of sources might benefit the discussion.

4. Does each paper support its claims and conclusions with explicit arguments or evidence?

Wang et al. (2023): The question how AI biases and explanations change the perceived fairness of a decision helped by AI is answered by showing that both a higher bias and the existence of explanations increase the perceived unfairness of AI enhanced decisions. This is shown by analysing the test results of the experiment statistically, showing that a bias in models as well as explanations introduce a bias in decisions based on the models.

Lai et al. (2023): Yes, the paper provides explicit arguments supported by evidence from the surveyed studies. The authors discuss trends and gaps in the literature, using data from their analysis to substantiate their conclusions and recommendations for future research. For each sub-group of literature, i.e., decision tasks, AI models, and evaluation metrics, different positions are presented. Decision tasks are grouped by the domain in which they can be applied, listing several tass where a human

decision can be enhanced by using help from an AI system. They do not only stress opportunities, but also keep an eye on risks when identifying trends and stress tasks where human intuition is still necessary. As gap, e.g., they find that the choice of tasks suitable for receiving help by AI, often depends on data set availabilities rather than applicable use cases, which would be a more reasonable focus. Regarding AI models and AI assistant elements, it is presented how predictive AI models, together with information about how the predictions are derived and information about the models themselves can supplement the decisions of humans. Here, they also give an extensive overview of different gaps and weaknesses of the current body of literature, showing room for improvement. Lastly, considering the evaluation of human-AI decision making, they elaborate different ways of evaluating the strengths of human-AI interactions in decision processes. Here, especially the necessity of working towards one common metric is stressed as this helps unify the body of research. Overall, with this classification of three different research tasks within the overarching topic, and the presentation of the status together with strengths and weaknesses of the current research in these fields, they are able to get across what is important and what is currently lacking, this way generally reaching the objective of their research.

5. How would you enhance the work/paper?

Wang et al. (2023): As already explained, the study is well-conducted. However, its generalisation is limited due to the laboratory setting in which it takes place. Conducting a real case study which analyses the behaviour of human beings in their natural lives outside of a laboratory environment might make more sense. However, also here I expect problems of a limited generalisation to arise due to humans' change in behaviour due to being observed. Still, this change of behaviour might be

smaller outside of a laboratory environment that inside of it. Furthermore, it might make sense to include multiple forms of human decision-making. This way, the study could profit from expanding its scope on also including other aspects of decision-making and showing whether the found effect that a higher bias and explanations increase the perceived unfairness of AI-supported decisions prevails.

Lai et al. (2023): This paper could be enhanced by providing a more detailed analysis of the impact of different study design choices on the outcomes of human-AI decision-making. Including meta-analyses or statistical evaluations of the effectiveness of various AI assistance elements could offer deeper insights. Furthermore, discussing the ethical considerations in human-AI decision-making studies would add a valuable dimension to the survey. While the study manages to present different study designs, their impact is only weakly covered and when it is covered the focus is on evaluating the concepts theoretically. Answering how these concepts are expected to positively influence employees' daily lives would be beneficial. This could be achieved by, e.g., investing in field studies in the work space.

List of References:

Wang, X., Liang, C., Yin, M. (2023) 'The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets', *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3076-3084. Available at: https://doi.org/10.24963/ijcai.2023/343

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V. and Tan, C. (2023) 'Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies', *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1369-1385. Available at: https://dl.acm.org/doi/10.1145/3593013.3594087