SIGHCI 2021 Proceedings

Special Interest Group on Human-Computer Interaction

12-12-2021

# Me, You and AI – Managing Human AI Collaboration in Computer Aided Intelligent Diagnosis

Maximilian Grüning

Manuel Trenz

# Me, You and AI – Managing Human AI Collaboration in Computer Aided Intelligent Diagnosis

**Maximilian Gruening**
University of Goettingen
maximilian.gruening@uni-goettingen.de

**Manuel Trenz**
University of Goettingen
trenz@uni-goettingen.de

## ABSTRACT

This research-in-progress paper focuses on examining configurations of collaboration between physicians and AI in decision making. From prior literature, we know that complex decisions in hospitals are the result of a collaborative decision-making process among physicians in a team. However, research from an information systems (IS) perspective in this area has so far focused on individual's interactions with AI, while collaboration in decision-making for complex clinical cases reflects common practice also in technologically supported environments. Therefore, we aim to shed light on the question "which configuration of human AI collaboration in decision making is most recommendable for AI-enabled systems?" We plan to conduct a scenario-based experiment to investigate accuracy, speed, and satisfaction with various configurations of physician AI collaboration in the context of computer-aided intelligent diagnosis (CAID) systems. Our primary contribution will be a multidimensional evaluation of selected collaboration configurations aimed at improving healthcare with technology.

## Keywords

Artificial Intelligence, Computer-aided intelligent diagnosis, Human AI collaboration, Healthcare.

## INTRODUCTION

When we watch TV shows about physicians, we ideate those humans as all-rounders in white coats acting as a solitary detective in hospitals. However, the general circumstances of their real-world have changed in recent years as high pressure in accuracy, overtime and resource scarcity shape daily business in clinical practice. Based on this observation, we see a dilemma of decision-making in knowledge intense situations (Jussupow et al., 2021). On the one hand, physicians need to deliver fast results in consultation with other experienced peers (Mirbabaie et al., 2021). On the other hand, those results should be accurate and justified by viewing and evaluating growing amounts of data. This dilemma, leads to higher pressure and complexity in decision-making for diagnosis and making the right choice in a melting pot of imaginable diseases.

As part of the solution to these circumstances, AI has the role to secure both, accuracy and speed, by using algorithms to detect and evaluate anomalies in visual qualities or health data sets (Lai et al., 2021). Although AI is sophisticated at diagnosing faster and more efficiently than physicians, there are ethical considerations that support placing the final decision about a patient's treatment in the hands of physicians. Therefore, the active interaction between physicians and AI in the diagnosis process could shape a new form of work routine in clinical practice. (Jussupow et al., 2021). However, we are unsure what form of collaboration with AI is required in complex situations. While recent research shows that collaborative decision-making with AI is different (Fügener et al., 2021), we do not know what this notion means for the clinical practice where we observe little evidence on the evaluation of human-AI collaboration (Mirbabaie et al., 2021).

To the best of our knowledge this is the first research that takes an IS perspective on configurations and decision-making patterns of human-AI collaboration in healthcare. To extend the existing body of knowledge, we tackle the topic of human-AI collaboration with the following research question: "Which configuration of collaboration between human-AI collaboration in decision making is most recommendable for AI-enabled systems?" To answer this question, we will observe actual physicians' diagnosis processes in a scenario-based experiment aiming to shed light on how physicians' accuracy, speed, and satisfaction manifest depending on the configuration of the collaboration with other physicians and/or an AI-based CAID system.

The research in progress paper is structured as follows. First, we explain the foundations of a radiological procedure and review the most current advances in the field of managing AI-enabled systems in a healthcare context. Areas such as radiology which already uses big amounts of data to detect diseases are pioneering in digitizing healthcare (Buck et al., 2021). Second, we summarize the foundations of human-AI collaboration. The subsequent section presents our experiment explaining the effects of AI advice on physicians' accuracy in different configurations, the speed of each diagnosis in those configurations, and possible manifestations of satisfaction of each physician. Following the experiment, we present the possible contributions of our research.

## RESEARCH CONTEXT: RADIOLOGICAL PROCEDURE

In the clinical practice of a university hospital, patients are directed to the radiology department via two streams. Either there are acute examinations from the emergency department or planned examinations from the ward. Both streams will be forwarded to the radiologists on duty. After the radiologists have taken the request, they transport the patient to a computed tomography (CT) scan and start with preliminary checkups. They collect patient data to calibrate the CT system, for instance to adjust the density of the measurement or choose the injection for the contrast medium. Subsequently they start with a CT scan. The Picture Archiving and Communication System (PACS) creates CT images of the patient and transfers them back to the system of the radiologists. Then the physicians begin the radiological evaluation and diagnosis. The radiologists have to examine the CT images and write a report. In the occasion of an unclear case, another physician will be consulted to initiate further measures. But more often in complex cases, the radiologist exchanges information with other experienced peers from the radiology department to solve the clinical problem by active collaboration. This occasion usually involves a team of physicians. At this part of the radiological diagnosis process, we see a possible interface for examining AI in different configurations of collaboration for decision making. Figure 1 illustrates the radiological procedure.

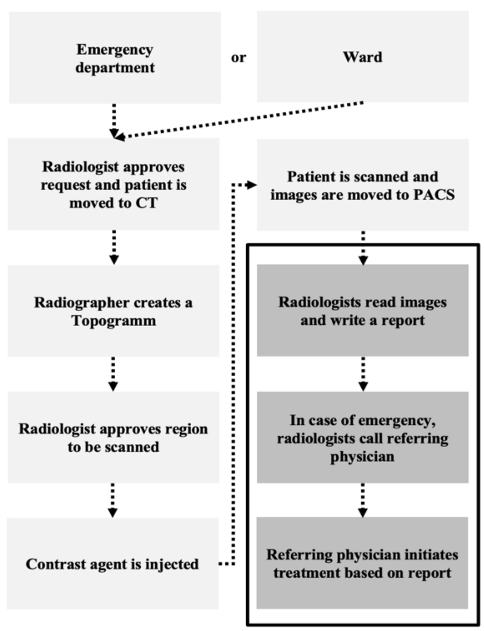


**Figure 1. Radiological Procedure**

## THEORETICAL BACKGROUND

The following section introduces our theoretical background drawing on two different layers: managing AI-enabled systems and human-AI collaboration. While these research fields have distinctive characteristics, they complement each other in highlighting the central role of humans and the new avenues of collaboration between humans and AI.

### Managing AI-enabled Systems

AI affects how we will make decisions in the future. In progressively complex situations, AI embodies a game-changer role to satisfy the demand for quicker and validated decisions by making large amounts of data accessible, usable and utilizable. Moreover, AI can take on a superior role in interaction with humans. Precisely AI can outperform humans (Shen et al., 2019) or outperform human crowds (Fu et al., 2021). Seeing AI as a frontier expands our horizon of understanding by showing that AI is perceived not only as a phenomenon, but rather as a moving target of evolving phenomena (Berente et al., 2021). Following Berente et al. (2021, p. 12), we define AI "as the frontier of computational advancements that references human intelligence in addressing ever more complex decision-making problems." This frontier embeds two dimensions, performance and scope. Performance describes the "ever-improving execution of tasks to which AI is applied while scope describes the "ever-expanding range of contexts to which AI is applied" (Berente et al., 2021, p. 12). In sum, managing AI also means making key decisions with the co-operation of AI, which we further describe in AI-enabled systems.

AI-enabled systems use an innovative approach for novel developments and applications (Rzepka & Berger, 2018). As AI-enabled systems address complex decision-making with references to human intelligence, we realize the spawn of CAID systems in clinical practice as a decent starting point to further research. CAID systems make more data available and accomplish tasks that were previously regarded as uniquely human (Jussupow et al., 2021). They offer a second diagnostic opinion for physicians' medical decisions and enable a correction of the previous diagnosis (Cheng et al., 2016). CAID systems have outperformed expert physicians in various contexts like diabetes or cancer (Shen et al., 2019). These systems help physicians prone to decision errors with different levels of experience, specializations or resilience (Shen et al., 2019).

### Human AI Collaboration

Human-AI research is an emerging topic in IS. Following Lai et al. (2021, p. 390) we define collaboration as "an evolving, interactive process whereby two or more parties actively and reciprocally engage in joint activities aimed at achieving one or more shared goals." In detail, human-AI collaboration refers to the collaboration between single or multiple humans and AI systems (Lai et al., 2021, p. 390). In the current research, we find different streams of collaboration between humans and AI. Most discoveries reveal the need for human-AI collaboration solutions and recommend how these solutions should be implemented in the organization (Davenport & Ronanki, 2018). We decide to focus on the concepts of AI and the configurations of collaboration between humans and AI in the healthcare context. While in general algorithm appreciation is preferred over human judgment (Logg et al., 2019), recent findings postulate that physicians tend to prefer the advice of a human expert in complex situations while they interact with an AI system (Jussupow et al., 2021). Notably Jussupow et al. (2021) focuse on the use of AI in clinical practice and identify decision patterns for individual physicians in the interaction with AI. Particularly the emergent configurations of collaboration between humans and AI for higher performance (Fügener et al., 2021) allow new opportunities related to task substitution, task augmentation or task assemblage. Thus, we plan to investigate the configuration of collaboration on the individual and team level of physicians while they analyze multiple CT images and find out which configuration in decision making is most advisable for their clinical practice.

For receiving an overview of the state of the art research on human-AI collaboration we conducted a structured literature review in IS, business and management literature (Brocke et al., 2009; Okoli & Schabram, 2010; Webster & Watson, 2002). In the structured literature review we follow five main phases of definition of scope, concept integration, literature search, analysis and forming the research agenda (Brocke et al., 2009) which leads to further findings. We find that configurations of human-AI collaboration could be classified as human-centered or AI-centered. We observe human-centered as the human providing input to the AI when needed further known as humans-in-the-loop-of AI (Boukhelifa et al., 2020; Holzinger et al., 2019) Other authors argue that AI systems should most likely be used as a tool that integrates into the process of human work (Davenport & Ronanki, 2018; Kocaballi et al., 2020). In our context, we understand the interaction with AI-enabled systems as an incorporation of meaningful interaction with our CAID system. Therefore, we understand this CAID systems as a tool wielded through human interaction that helps to augment decisions.

Lastly we define hybrid intelligence (HI) as "the ability to achieve complex goals by combining human and AI, thereby reaching superior results to those each of them could have accomplished separately, and continuously improve by learning from each other" (Dellermann et al., 2019, p. 640). We receive the collaboration mechanism between humans and AI to solve a task as key for our teamwork research as illustrated in Figure 2 (Hemmer et al., 2021).

### Theoretical Model

The improvement of the diagnostic outcome when a single physician use AI as a tool in a clinical context is well documented in interdisciplinary studies. However, it is unclear which configuration of collaboration between humans as a team and AI leads to the best possible outcome in the routine of clinical practice. We argue that with the increasing utility of AI in a team, the quality of diagnosis increases by the following:

*Hypothesis 1: Human accuracy in diagnosis differs between collaboration such that a team of humans and AI > one human and AI > a team of humans > one human*

*Hypothesis 2: Human speed in diagnosis differs between collaboration such that one human and AI > a team of humans and AI > a team of humans > one human*

*Hypothesis 3: Human satisfaction in diagnosis differs between collaboration such that a team of humans > a team of humans and AI > one human and AI > one human*

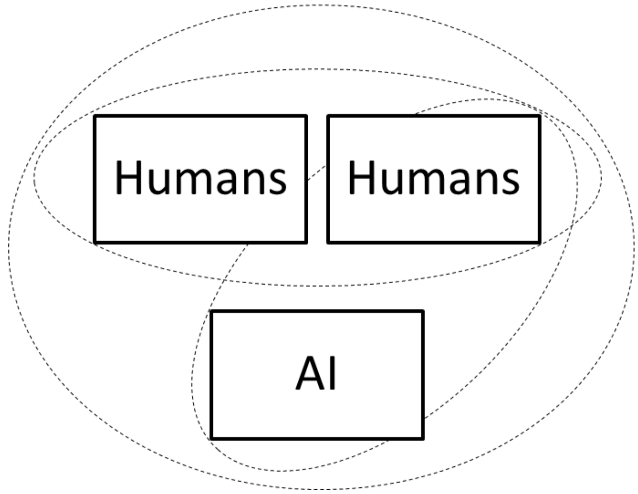


**Figure 2. Humans-in-the-Loop**

Next, we present our experimental design to study configurations of collaboration between humans and AI in decision making and to test the hypotheses derived from our theoretical model.

## STUDY DESIGN

To empirically test our hypotheses, we plan to conduct a 2x2 set of experimental studies with physicians and AI. The physicians have different levels of experience and will diagnose real patient CT images under real-world conditions in a selected location in a hospital. The patient CT images were selected from previous radiological examinations by two senior physicians, that do not participate in the experiment. Further, we provide an overview of our experimental design and the configuration of each experimental group in Table 1.

| | Individual Level (1) | Team Level (2) |
|---|---|---|
| Without AI (A) | | |
| With AI (B) | | |

**Table 1. Experimental Groups**

We follow a positivist approach and describe our data collection process, the instrument development process and the sampling process in the following. The design consists of two steps. First, we start the main scenario-based experiment in AI-enabled diagnosis. Second, we send a questionnaire to the participants after the main experiment.

### Participants

We invite a broad variety of physicians from two university hospitals. Overall, we plan to have 100 participants in our scenario-based experiment. All physicians will participate voluntarily and take part from their local place of work.

### CAID System

For our main experiment we use a newly developed CAID system from our research project partner that predicts pulmonary embolism (PE) diagnosis from visual characteristics. We choose PE because it is one of the most common medical cases (Li et al., 2021). In particular, PE is characterized by a high mortality rate, high morbidity and is often overlooked in complex occurrence in clinical practice (Li et al., 2021). To shed light on undetected occurrences, collaboration with peers and a data-driven CAID system incorporated in a CT can lay the foundation for the smart discovery of diseases. We illustrate the interface of our CAID system in Figure 3 and choose the context of diagnosis in radiology for one main reason: CT image detection is a common practice to detect specific diseases in radiology. In detail, the AI of the CAID consists of a three-staged model predicting whether a patient has PE and a location label. The first stage is a preprocessing step in which the lung region is extracted. This ensures further models to appeal precise and makes the problem easier for further stages. The second stage is a Convolutional Neural Network (CNN) is trained with the PE labels "positive" or "negative" to extract the relevant features of each 2D image.  This features extractor can be trained on data from multiple medical centers and ensures a more robust features are extracted. Finally, the AI combines the features from the previous stage with a transformer model overall 2D slices. This allows the model to gain context over the whole 3D CT image data of the patient and predict location labels.

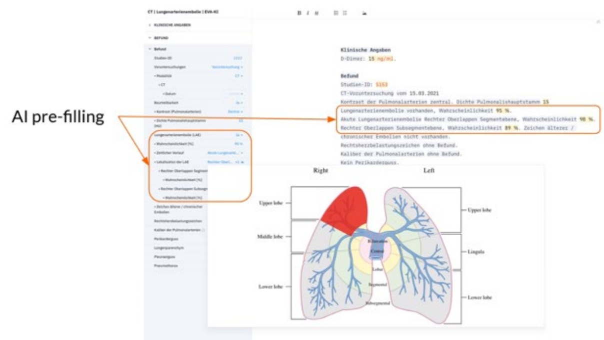


**Figure 3. Interface of the CAID System**

### Experimental Setup

After the introduction, we divide the physicians into four randomized groups. Each individual has the same probability to be assigned to one of the groups. By this allocation, possible confounding variables are minimized. From related research we know that experiment-based research articles show the AI-supported systems to physicians in a medical context with high ambiguity (e.g., breast cancer detection). Further most research papers apply qualitative or quantitative research methods. Some authors also combine both e.g., physicians have interviews and questionnaires before and after the experiment to assess the familiarity with AI (Jussupow et al., 2021).

### Measurements

We evaluate the different configurations using performance and perceptual data. To measure the performance data of each group, we collect data on the accuracy and speed of each diagnosis session. In addition, we measure satisfaction and who actually influenced the final decision for the diagnosis (physician or AI) in three treatment groups and one control group as part of perceptual data. Questions about the purpose of the CAID system and whether they received support in diagnosis from the CAID will serve as manipulation checks. Lastly, we will ask questions about demographics like level of expertise, work experience, age, number of clinical cases per anno, to account for related factors.

## INTENDED CONTRIBUTIONS

We plan to contribute to research on human-AI and the design of AI-enabled decision support configurations. First, while recent research indicates that the interaction of humans with AI has the potential to increase performance measures, we shed light on the potential of alternative human-AI configurations (Fügener et al., 2021) such as collaborative environments in clinical practice of the radiological procedure. Thereby, we expand our understanding of decision patterns in collaboration between physicians and AI in medical diagnosis to a team environment (Jussupow et al., 2021). Second, we complement quantitative evaluations of human-AI collaborations by qualitative indicators like satisfaction. Our multidimensional evaluation in a realistic work setting will provide a fine-grained understanding of not only how effective certain configurations may be, but also which types of collaborations individuals will support in attractive future work environments where humans and intelligent technologies work together.

## REFERENCES

Berente, N., Gu, B., Recker, J., & Santanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Boukhelifa, N., Bezerianos, A., Chang, R., Collins, C., Drucker, S., Endert, A., & Hullman, J. (2020). Challenges in Evaluating Interactive Visual Machine Learning Systems. *IEEE Computer Graphics and Applications*, 88–96. https://doi.org/10.1109/MCG.2020.3017064

Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. *Proceedings of the 17th European Conference on Information Systems (ECIS 2009)*.

Buck, C., Hennrich, J., & Kauffmann, A. L. (2021). Artificial Intelligence in Radiology – A Qualitative Study on Medical Imaging Specialists' Perspectives. *Proceedings of the 42th International Conference on Information Systems* (ICIS 2021).

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., & Chen, C.-M. (2016). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. *Scientific Reports*, *6*(1), 24454. https://doi.org/10.1038/srep24454

Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence for the Real World. *Harvard Business Review*, *96*(1), 108-116.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, Lending, Machine, and Bias. *Information Systems Research*, *32*(1), 72–92. https://doi.org/10.1287/isre.2020.0990

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, *45*(3), 1527–1556. https://doi.org/10.25300/MISQ/2021/16553

Hemmer, P., Vössing, M., Schemmer, M., & Kühl, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *Proceedings of the 25th Pacific Asia Conference on Information Systems* (PACIS 2021).

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., & Palade, V. (2019). Interactive machine learning: Experimental evidence for the human in the algorithmic loop: A case study on Ant Colony Optimization. *Applied Intelligence*, *49*(7), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Kocaballi, A. B., Ijaz, K., Laranjo, L., Quiroz, J. C., Rezazadegan, D., Tong, H. L., Willcock, S., Berkovsky, S., & Coiera, E. (2020). Envisioning an artificial intelligence documentation assistant for future primary care consultations: A co-design study with general practitioners. *Journal of the American Medical Informatics Association*, *27*(11), 1695–1704. https://doi.org/10.1093/jamia/ocaa131

Lai, Y., Kankanhalli, A., & Ong, D. (2021). Human-AI Collaboration in Healthcare: A Review and Research Agenda. *Proceedings of the 54th* Hawaii International Conference on System Science (HICSS 2021).

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mirbabaie, M., Stieglitz, S., & Frick, N. R. J. (2021). Hybrid intelligence in hospitals: Towards a research agenda for collaboration. *Electronic Markets*, *31*(2), 365–387. https://doi.org/10.1007/s12525-021-00457-4

Okoli, C., & Schabram, K. (2010). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1954824

Rzepka, C., & Berger, B. (2018). User Interaction with AI-enabled Systems: A Systematic Review of IS Research. *Proceedings of the 39th International Conference on Information Systems (ICIS 2018)*.

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P.-H., & Ming, W.-K. (2019). Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR Medical Informatics*, *7*(3), e10010. https://doi.org/10.2196/10010

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii.