AUGMENTING MEDICAL DIAGNOSIS DECISIONS? AN INVESTIGATION INTO PHYSICIANS' DECISION MAKING PROCESS WITH ARTIFICIAL INTELLIGENCE

Jussupow, Ekaterina - University of Mannheim, Germany, jussupow@uni-mannheim.de

Spohrer, Kai – University of Mannheim, Germany, spohrer@uni-mannheim.de

Heinzl, Armin – University of Mannheim, Germany, heinzl@uni-mannheim.de

Gawlitza, Joshua – Saarland University Medical Center, Germany, Joshua.Gawlitza@uks.eu

*** Paper accepted for publication in Information Systems Research.**

ABSTRACT

Systems based on artificial intelligence (AI) increasingly support physicians in diagnostic decisions. Compared to rule-based systems, however, these systems are less transparent and their errors less predictable. Much research currently aims to improve AI technologies and debates their societal implications. Surprisingly little effort is spent on understanding the cognitive challenges of decision augmentation with AI-based systems although these systems make it more difficult for decision makers to evaluate the correctness of system advice and to decide whether to reject or accept it. As little is known about the cognitive mechanisms that underlie such evaluations, we take an inductive approach to understand how AI advice influences physicians' decision making process. We conducted experiments with a total of 68 novice and 12 experienced physicians who diagnosed patient cases with an AI-based system that provided both correct and incorrect advice. Based on qualitative data from think-aloud protocols, interviews, and questionnaires, we elicit five decision making patterns and develop a process model of medical diagnosis decision augmentation with AI advice. We show that physicians use distinct metacognitions to monitor and control their reasoning while assessing AI advice. These metacognitions determine whether physicians are able to reap the full benefits of AI or not. Specifically, wrong diagnostic decisions often result from shortcomings in utilizing metacognitions related to decision makers' own reasoning (self-monitoring) and metacognitions related to the AI-based system (system-monitoring). As a result, physicians fall for decisions based on beliefs rather than actual data or engage in unsuitably superficial information search. Our findings provide a first perspective on the metacognitive mechanisms that decision makers use to evaluate system advice. Overall, our study sheds light on an overlooked facet of decision augmentation with AI, namely the crucial role of human actors in compensating for technological errors.

Keywords: Decision Making, Artificial Intelligence, Decision Support, Metacognition, Healthcare, Cognition, Advice Taking

INTRODUCTION

The rapid development of technologies based on artificial intelligence[1] (AI) has shattered the notion of physicians as the sole decision makers in clinical practice. Advances in machine learning with increased data availability have spawned *computer aided intelligent diagnosis* (CAID) systems that accomplish tasks that were previously regarded as uniquely human (Mayo and Leung 2018). In fact, AI-based systems have begun to outperform expert physicians in diagnosing diseases such as diabetes, cancer, and stroke (e.g., Shen et al. 2019). Yet, the complexity and consequences of medical decisions make it unlikely that CAID systems will replace physicians in the diagnostic process (Jha and Topol 2016). Instead, CAID systems promise to augment physicians' medical decisions by providing a second diagnostic opinion and by offering an opportunity to revise preliminary diagnostic assessments if necessary (Cheng et al. 2016). Thus, CAID systems should be beneficial for those physicians who are most susceptible to diagnostic errors, including novice physicians with limited experience, physicians with different areas of specialization, and those working on complex cases under high cognitive load and time pressure (Shen et al. 2019). Overall, the combined assessment of CAID systems and physicians should result in lower error rates than decisions made by physicians or CAID systems alone (Cheng et al. 2016, Mayo and Leung 2018).

However, including AI advice in clinical practice creates new challenges. Although CAID systems have reached high levels of accuracy, they are not without errors. Unlike rule-based clinical decision support systems, AI-based systems are based on statistical data patterns rather than explicit human expertise (Appendix A provides a detailed comparison). Thus, distorted data can lead to errors that are unpredictable for both CAID system developers and physicians (Rahwan et al. 2019). For example, AI-based systems can

[1] AI refers to the ability of a computer to accomplish tasks commonly associated with intelligent beings, i.e., intellectual processes that are characteristic for humans such as reasoning, generalizing, or learning from experience (Russell and Norvig 2010).

display racist and sexist decision schemes because of distortions in training data (Kirkpatrick 2016). At the same time, identifying incorrect assessments of CAID systems is inherently difficult due to their technological properties. For example, many CAID systems in radiology rely on deep learning algorithms to offer diagnostic advice based on imaging data (Jiang et al. 2017). These systems' inference logic is necessarily less transparent than the logic of traditional rule-based systems. CAID systems typically provide solely the results of their analysis, while the reasoning remains a black box (Fazal et al. 2018). Thus, it is difficult, yet critical, that physicians supervise CAID systems and do not follow AI advice without scrutiny.

Whereas much research currently focuses on how to develop more accurate and transparent AI-based systems to counter these challenges (Rai 2020), too little effort is spent on understanding the cognitive challenges of decision makers that hinder successful decision augmentation with AI advice (Burton et al. 2020). On the one hand, physicians often ignore system advice (Liberati et al. 2017) and, thus, fail to benefit from the second opinion that increasingly powerful CAID systems provide. On the other hand, decision makers often fail to detect incorrect system advice and are misled by it. In prior studies, experienced physicians failed to overrule incorrect system advice in between 33% and 48% of all cases when examining mammograms (Alberdi et al. 2004). Similarly, physicians interpreted electrocardiograms more accurately with the help of correct system advice but dropped far below their unsupported accuracy levels when they received occasionally incorrect system advice (Tsai et al. 2003). It is therefore necessary to comprehend and mitigate physicians' cognitive challenges with CAID systems that can provide correct as well as incorrect advice. Failure to detect incorrect AI advice may otherwise result in wrong diagnoses and numerous medical errors (Tsai et al. 2003, Alberdi et al. 2004, Goddard et al. 2012).

From a theoretical perspective, little is known about the cognitive mechanisms that allow decision makers to evaluate the correctness of system advice and decide whether to reject or accept it. In fact, most prior work in the Information Systems (IS) literature has been assuming that system advice is correct and beneficial (e.g., Davern et al. 2012, Arnott and Pervan 2014). In doing so, the literature largely neglected the cognitive challenges that incorrect system advice poses to decision makers and why they often fail to reject it (Schultze et al. 2017, Fiedler et al. 2019). Moreover, prior work provided a theoretically fragmented

picture of the cognitive challenges involved in augmenting decisions with system advice. On the one hand, prior work has suggested that decision makers need to deliberately process system advice in order to integrate it successfully into their decision making (e.g., Heart et al. 2011). On the other hand, prior work showed that decision makers often fail to do so because they apply heuristic reasoning and do not switch from quick, superficial assessments to deliberate, in-depth reasoning when necessary (e.g., Adomavicius et al. 2013). Although crucial for decision augmentation, research has not yet understood the mechanisms through which decision makers successfully balance deliberate, in-depth reasoning with quick, superficial assessments of system advice (Ferratt et al. 2018).

We therefore turn to research in psychology which theorized that decision makers dynamically balance quick, heuristic reasoning and deliberate, systematic reasoning by means of metacognitions (Fiedler et al. 2018, 2019). Metacognitions are second-order cognitive processes that monitor and control human decision making (Ackerman and Thompson 2017). Throughout a decision making process, different types of metacognitions help decision makers to reflect on their own reasoning, to decide what information to consider and how to account for it. Thus, metacognitions also influence whether and how decision makers consider provided advice (Fiedler et al. 2019). Given that decision makers treat system advice very differently from human advice (Dietvorst et al. 2015, Logg et al. 2019, Longoni et al. 2019), it is still unclear which types of metacognitions allow decision makers to successfully include system advice in their decision making process. Scholars in psychology recently suggested that it is particularly important to better understand which metacognitions allow decision makers to cope with potentially incorrect advice (Fiedler et al. 2019). In order to understand successful decision augmentation with AI advice, we consequently examine the role of metacognitions throughout physicians' decision making process with CAID systems. We aim to answer two research questions:

1) How does diagnostic AI advice influence physicians' decision making process?
2) Which metacognitions do physicians use to decide whether to follow or to discard AI advice?

FOUNDATIONS OF DECISION AUGMENTATION AND METACOGNITIONS

In the following, we first conceptualize the challenges of decision augmentation with CAID systems and then synthesize prior work that addressed the underlying decision making mechanisms. Finally, we outline the theoretical background of metacognitions and their role in human decision making.

Decision augmentation of medical decisions with system advice

We refer to *decision augmentation* as the process through which systems provide advice[2] to decision makers that enhances resulting decision outcomes. Decision makers consider, evaluate, and balance the advice against their own assessment, and derive decisions that can be measured in terms of accuracy and quality. Decision augmentation with CAID systems can be conceptualized as physicians' evaluation of their own diagnostic assessment against provided AI advice (see Table 1). In decisions in which correct AI advice confirms a physician's correct assessment (Confirmation I), physicians are reinforced in their assessment. Thus, they are likely to retain their initial assessment. Similarly, physicians and CAID systems can both come to an incorrect assessment (Confirmation II) which likely results in medical errors. If correct AI advice conflicts with a physician's incorrect assessment (Disconfirmation I), successful decision augmentation would require that physicians change their assessment and accept the AI advice. By contrast, if incorrect AI advice conflicts with a physician's correct assessment (Disconfirmation II), successful decision augmentation would require that they retain their initial assessment and discard the advice. Consequently, successful decision augmentation requires different behavioral responses from physicians depending on the correctness of the AI advice they receive. Recognizing and executing the desirable behaviors poses, however, significant cognitive challenges.

Table 1. Conceptualizing decision augmentation of medical decisions

		Physician's initial assessment	
		Correct	**Incorrect**
CAID system	**Correct**	Confirmation I: correct assessment is reinforced	Disconfirmation I: correct system advice Desirable behavior: change own assessment and accept AI advice

[2] In analogy with human advice, we define AI advice (Bonaccio and Dalal 2006) as a recommendation or suggestion by an AI-based system to a decision maker for a specific decision task.

advice	**Incorrect**	Disconfirmation II: incorrect system advice Desirable behavior: retain own assessment and discard AI advice	Confirmation II: worst-case scenario as decision makers do not detect problem

Prior work on cognitive challenges of decision augmentation with system advice

Our study is informed by three lines of research (Table 2) which conceptualize cognitive challenges in decision augmentation as revolving around heuristic or systematic reasoning processes. Systematic reasoning refers to the detailed, controlled, and deliberate cognitive processing of system advice, which is relatively slow and effortful (Evans and Stanovich 2013, Ferratt et al. 2018). Heuristic reasoning refers to mental shortcuts that constitute quick and automatic cognitive responses to a trigger without effortful processing of details (Tversky and Kahneman 1974, Kahneman 2011). Heuristic reasoning, thus, provides a fast default response to utilize or not utilize system advice (cf. Ferratt et al. 2018).

A first, traditional IS research stream has intensively focused on understanding and stimulating the utilization of system advice. This line of work has investigated challenges of Confirmation I and Disconfirmation I and assumed that system advice is correct and beneficial. It demonstrates many benefits of system advice including improved adherence to medical guidelines, decreased medical errors, and increased medical decision quality (Jaspers et al. 2011). From a theoretical point of view, this research stream holds that the success of decision augmentation is driven by how deeply decision makers integrate system advice into their decision making. They need to engage in systematic reasoning to reflect on the advice content and its quality (Xiao and Benbasat 2007, Heart et al. 2011, Arnott and Pervan 2014). This research stream also aims at facilitating advice utilization by designing systems that foster decision makers' systematic reasoning, for example by providing explanations (Arnold et al. 2006). In doing so, this research stream has elaborated primarily on how systematic reasoning supports decision augmentation but largely ignored cognitive decision processes that hinder optimal utilization of system advice.

Table 2. Overview of three literature streams that address challenges of decision augmentation

Research stream	Advice source	Augmentation challenge and assumptions	Identified cognitive challenges	Underlying reasoning mode	Exemplary paper

Stream 1: Decision augmentation with correct system advice	Decision support & recom-mender systems	Confirmation I & Disconfirmation I – Advice is correct and beneficial	Decision makers insufficiently integrate system advice into their decision making	Systematic reasoning necessary to integrate system advice into their decision making	Arnold et al. 2006, Nissen and Segupta 2006, Xiao and Benbasat 2007, Tan, Teo and Benbasat 2010, Heart et al. 2011, Davern et al. 2012, Arnott and Pervan 2014
Stream 2: Human biases in evaluating system advice	Algorithm advice	Disconfirmation I – Advice is correct and beneficial but imperfect	Decision makers have inherent biases against or towards algorithms	Heuristic reasoning hinders optimal utilization of advice	Dietvorst et al. 2015, 2018, Logg et al. 2019, Longoni 2019
	Decision support & recom-mender systems	Confirmation II – Advice can be deceptive	Decision makers cannot compensate for biasing influence of the system		Xiao and Benbasat 2011, 2015, 2018, Adomavicius et al. 2013, Elkins et al. 2013
Stream 3: Dealing with incorrect advice	Performa-tive systems	Confirmation II – Advice can be incorrect	Decision makers lose vigilance in information processing	Decision makers fail to dynamically switch from heuristic to systematic reasoning	Tsai et al. 2003, Parasuraman and Manzey 2010, Goddard et al. 2012, Endsley 2017
	Human advisor	Disconfirmation II – Advice can be incorrect	Decision makers are unable to engage in systematic reasoning		Schultze et al. 2017, Fiedler et al. 2019

Notes. These categories describe different research streams and are not mutually exclusive. Algorithms are considered to be the outcome of statistical systems which are often more accurate than human decision makers (see Dietvorst et al. 2015). Recommender systems provide recommendations, often to consumers about preferences. Performative systems are highly automated systems in which decision makers monitor system performance, e.g., flight assistance systems (Nissen and Segupta 2006); they are mainly considered by human factors literature.

A second research stream in IS has suggested that distinct heuristic processes of human decision makers prevent adequate advice utilization (Xiao and Benbasat 2011, 2015, Dietvorst et al. 2015, 2018). For instance, decision makers who see themselves as experts have a tendency to be overly confident in their own assessments and often reject system advice (Elkins et al. 2013). Moreover, many decision makers unduly reject system advice if they realize that a system is imperfect (Dietvorst et al. 2015, 2018). Conversely, inexperienced decision makers often overestimate the correctness of systems compared to human advisors (Logg et al. 2019). Furthermore, system advice can anchor decision makers (Adomavicius et al. 2013) so that they are misled or deceived by unbeneficial system advice (Xiao and Benbasat 2011, 2015, 2018). From a theoretical point of view, this line of work suggests that heuristic processes hinder a sophisticated verification and utilization of system advice. On the one hand, inherent biases against systems are argued to explain why individuals fail to accept beneficial advice (Disconfirmation I). On the other hand, inherent biases toward systems are argued to explain why decision makers accept unbeneficial advice when it does not completely contradict their own assessment (Confirmation II). Although this line of work elaborated heuristic processes as reasons of inadequate advice utilization and proposed design features to

overcome single heuristic biases (e.g., additional instructions, explanations, and warning messages), it did not consider which deliberate cognitive processes actually help overcome such biases.

A final research stream addresses the challenges of Confirmation II and Disconfirmation II from a process perspective but has received little attention in the IS literature. This stream suggests that incorrect system advice is extremely problematic because decision makers often fail to detect occasional errors of highly accurate systems (Parasuraman and Manzey 2010, Goddard et al. 2012, Endsley 2017). In fact, decision makers are strongly influenced by incorrect advice, even if they know the advice is unreasonable (Schultze et al. 2017, Fiedler et al. 2019). From a theoretical point of view, these effects are explained as decision makers' failure to switch from heuristic reasoning to systematic reasoning when necessary. As such, decision makers lose vigilance for erroneous system advice if their systems usually perform well (Parasuraman and Manzey 2010, Endsley 2017). They fail to spot occasional errors because they rely increasingly on superficial heuristic assessments with insufficient systematic reasoning. Yet, the mechanisms through which decision makers balance their heuristic and systematic reasoning to process system advice have not been elaborated.

Overall, there are two important gaps in prior research. First, most prior work, especially in IS, has assumed that provided system advice is correct and beneficial. In doing so, it has largely neglected the cognitive challenges entailed in incorrect system advice, particularly the challenges of Disconfirmation II. Since even accurate AI-based systems might occasionally provide incorrect advice due to unnoticed distortions in training data, we need to understand the cognitive processes that enable decision makers to reject incorrect system advice while utilizing correct advice. Second, prior research has provided a theoretically fragmented picture of the cognitive challenges of decision augmentation with system advice. On the one hand, it has suggested that decision makers need systematic reasoning to benefit from system advice. On the other hand, it has demonstrated that decision makers are often unable to engage in systematic reasoning as they fail to switch from heuristic to systematic reasoning. Little attention, however, has been given to the underlying mechanisms that allow decision makers to balance quick heuristic assessments with effortful systematic assessments of system advice if necessary (Ferratt et al. 2018). In fact, research on

human advice recently suggested that elaborating these mechanisms is paramount for understanding how individuals can cope with the cognitive challenges posed by potentially incorrect advice (Fiedler et al. 2018, 2019). In line with these ideas, we deem it necessary for IS research to take a step back and examine more rigorously how decision makers control their reasoning processes to benefit from decision augmentation with AI advice.

Naturalistic decision making and metacognitions

We draw on a theoretical framework that does not focus on heuristic and systematic processes in isolation but scrutinizes the more general cognitive processes that allow decision makers to monitor and control their reasoning in complex decision processes. Naturalistic decision making (NDM) (Klein 2008, Klein 2015, see Table 4 on p. 15 for definitions of core concepts) asserts that decision makers first build an initial mental model (hereafter referred to as *frame*) which is based on an intuitive assessment of the decision task (Klein 2008; Klein 2015). Once they accumulate further cues from available data, decision makers are likely to retain their initial assessment if the data confirm their frame. If the data disconfirm their frame, decision makers engage in a sensemaking process with more effortful systematic reasoning to decide whether they should preserve or change their frame (Kahneman and Klein 2009). In particular, "[w]hen there are cues that an intuitive judgment could be wrong, [the decision maker replaces] intuition by careful reasoning" (Kahneman and Klein 2009, p. 519).

Balancing intuitive, heuristic and deliberate, in-depth reasoning activities is achieved through metacognitions (Ackerman and Thompson 2017). Decision makers use these metacognitions to monitor and control their own decision making process. *Metacognitive monitoring* captures the dynamic state of confidence regarding how well a decision is being performed (Ackerman and Thompson 2017). If decision makers are confident, they will act and are less likely to seek further information or additional cues like system advice (for an overview see Bonaccio and Dalal 2006). If they are not confident, they will hesitate, gather more information, change tracks, or look for other cues like system advice (Hausmann and Läge 2008, Wang and Du 2018), which makes them more likely to follow conflicting advice. Thus, metacognitive monitoring constitutes a sensing activity that allows decision makers to regulate the degree of systematic

reasoning and the amount of information sought (Ackerman 2014, Ackerman and Thompson 2017). In contrast, *metacognitive control* encompasses deliberate action to initiate, terminate, or change the input factors for decision making. By means of metacognitive control, decision makers influence the ongoing decision making process. The most prominent control functions are searching for additional information and ceasing the search for information, i.e., making a decision without further elaboration (Ackerman and Thompson 2017). Fiedler and colleagues (2018) argue that if decision makers fail to exert metacognitive control, they evaluate data in a biased way and often base their assessments on irrelevant information. A lack of metacognitive control leads to decision makers' inability to discard useless advice even if they recognize its limited usefulness (Fiedler et al. 2019). Overall, similar processes are likely to occur when physicians face incorrect AI advice.

In sum, metacognitions allow decision makers to monitor and control their reasoning activities throughout the decision making process. Notably, metacognitions have rarely been studied in the context of complex decisions with potentially incorrect advice (Fiedler et al. 2019), let alone system advice. Thus, it is currently unclear which types of metacognitions decision makers use to evaluate AI advice and how metacognitions lead decision makers to follow or reject the AI advice.

METHOD

To research how physicians apply metacognitions in their decision making process with AI advice, we chose an inductive approach in an experimental setting in which we manipulated the correctness of advice provided by a CAID system.

Research Design

We conducted a first controlled experiment with 47 novice physicians who had to make diagnostic decisions for patient cases based on radiological data and advice provided by a CAID system. For triangulation, we (1) compared our findings from the first experiment with a sample of 12 experienced radiologists, and (2) changed our experimental design and examined a second group of 21 novice physicians. The participants diagnosed patient cases with the support of a CAID system that provided randomized correct and incorrect mock-up AI advice. We collected qualitative data using think-aloud protocols,

interviews, and questionnaires to capture the participants' decision making processes. Through our data analysis, we elicited distinct decision making patterns that we used to develop a process model of medical diagnosis decision augmentation with AI advice.

The CAID system

We developed a CAID system that predicts pulmonary function values from a computed tomography (CT) scan with machine learning for diagnosing chronic obstructive pulmonary disease (COPD) (blinded). COPD is a chronic lung disease and constitutes the third leading cause of death worldwide (World Health Organization 2018). Current medical practice uses primarily pulmonary function tests to diagnose COPD. Using pulmonary function values predicted from CT image data is a novel approach that has the potential to help detect COPD in earlier stages (blinded). The CAID system included binary mock-up AI advice that read: "Based on the analysis of the above data, the AI recommends: COPD / NO COPD" (see Figure 1). In a short introductory video before the experiment, we explained to the participants how the CAID system works and how the predicted data correspond with the CT image. None of the participants questioned whether the mock-up AI advice was really based on AI methods. The introduction explicitly stated that the accuracy of the AI advice was 90%, which is comparable to the prevailing clinical practice (Jiang et al. 2017).

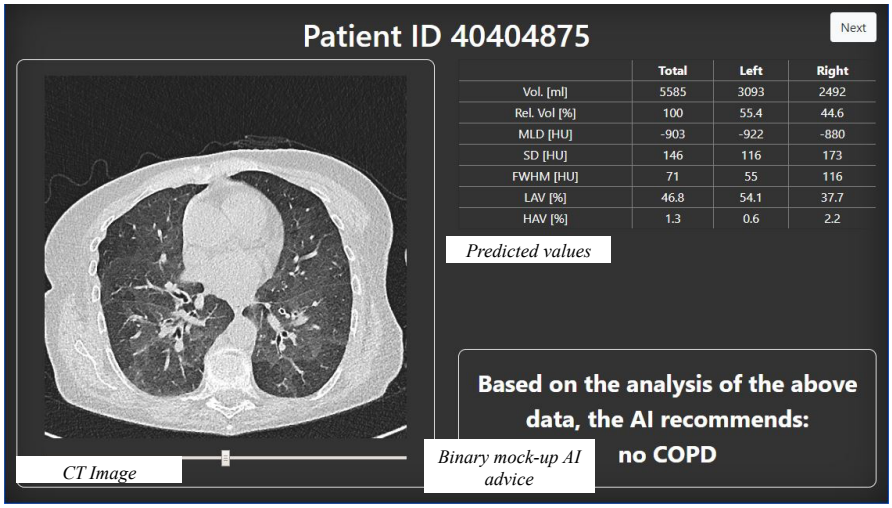

Figure 1. Interface of the CAID system. Predicted values are system predictions derived from a CT scan and refer to lung volume (Vol.), relative lung volume (Rel. Vol.), mean lung density (MLD), full-width-half-max (FWHM), low attenuation volume (LAV), and high attenuation volume (HAV)

Study procedure

To simulate critical clinical situations, we asked participants to conduct three diagnostic decisions based on the available information. We loaded three actual patient cases into the CAID system—two cases of patients suffering from COPD and one of a healthy patient. The three patient cases were selected and classified before the experiment by a radiologist and a pneumologist as unambiguous cases of patients who clearly suffered from COPD or not. We displayed the three patient cases in randomized order. Each participant completed three trials. In the first trial, we only provided the CT image without predicted values and without AI advice. This trial served as a control trial to understand the unsupported decision making process. In the second and third trials, participants received the full interface (see Figure 1). Each participant received one trial with correct AI advice and one with incorrect AI advice in random order. The participants could compare the AI advice with the predicted lung values in a table and with the CT image, which both indicated the correct diagnosis at all times. The AI advice was provided at the same time as all other data to allow participants to consider all information simultaneously and to simulate clinical situations in which radiologists must evaluate multiple sources of information simultaneously.

During each trial, we collected qualitative and quantitative data in multiple ways. First, we asked participants to complete the trial following the think-aloud method (van Someren et al. 1994). This method has successfully been applied in IS research to study decision making processes (e.g., Todd and Benbasat 1991, Li et al. 2017). We followed the guidelines of van Someren et al. (1994) and provided a cover story as well as a training task to familiarize participants with the procedure. The think-aloud protocols were recorded on video and audio. All verbal statements were transcribed. Second, at the beginning of the study and after each trial, we asked participants to complete a questionnaire indicating their decision confidence and satisfaction (survey items in online Appendix C). Finally, we conducted a short interview with the participants before debriefing. We asked them to reflect on their general attitude toward AI in healthcare and describe how they experienced the interaction with the AI advice during the study. The procedure, survey, and interview questions were pretested with a radiologist and two novice physicians. Table 3 provides an overview of the experimental procedure and the data collected. Participants required, on average, 2:21 minutes for the control trial (min = 0:41 minutes; max = 4:53 minutes) and 2:34 minutes for

the supported trials (min = 1:00 minute; max = 7:01 minutes). After each trial, participants had 40 seconds to rest and to prepare for the next trial. The total experiment lasted between 20 and 30 minutes.

Table 3. Study procedure and data collection

Step in the experiment	**Detailed procedure**	**Experimental manipulation**			**Data collection**
		Trial	**AI advice**	**Patient case**	
Introduction	- Welcome and consent - 3:47 min introduction video - Training in think-aloud task - Initial survey	None			- Demographic information - General expertise
Interaction with CAID system	- Task: Develop a diagnosis decision for the patient case (COPD or NO COPD) - Think-aloud - Brief survey after each trial	1	None	COPD/ NO COPD (randomized 2:1)	- Accuracy of the decision: correct and incorrect - Think-aloud protocols of the decision making process - Confidence and satisfaction ratings from survey - Observer notes
		2	Randomized correct and incorrect AI advice with COPD / NO COPD		
		3			
End of study	- Interview - Extensive debriefing	None			- Recorded, transcribed interviews

Participants

We selected novice physicians as participants for two primary reasons: First, since novice physicians are the target group for CAID support, CAID systems will have a stronger impact on their professional futures. Second, research has indicated that novice physicians are more likely to comply with CAID systems by adapting their diagnosis (Goddard et al. 2012). By choosing novice physicians, we experimentally overestimated the occurrence of medical diagnostic errors in order to identify different underlying reasons for these errors. Our first study included a mix of 47 participants consisting of 26 novice physicians without clinical experience (medical students with 4 years of medical training, on average) and 21 novice physicians with clinical experience (between 0.5 and 1.0 years of clinical experience).

Data triangulation

We used two additional data samples to triangulate our findings. First, we conducted a second data collection using a slightly varied interface but the same experimental procedure. In this experiment, the AI advice was provided after the display of the CT image and the predicted values (see online Appendix D for more details on the experimental set-up). This ensured that all participants made an explicit initial assessment before engaging with the AI advice and then compared their assessment with the AI advice. This reversed-order experiment involved a sample of 21 novice physicians, consisting of nine physicians

in their first clinical year (between 0.5 and 1.0 years of clinical experience) and 12 advanced medical students (4.5 years of medical training, on average). Second, we collected data from a sample of experienced radiologists (n = 12) with an average of 9.28 years of clinical experience. Of those, four received the original experimental design while the remaining eight received the reversed-order design. Triangulating our patterns and process model using experienced physicians helped us identify the mechanisms underlying the decision process, and to transfer our findings into clinical practice.

Data Analysis

Two researchers coded the data in an iterative approach of descriptive, axial, and theoretical coding (Strauss and Corbin 1990, Saldaña 2013). Table 4 depicts the descriptive codes including the predefined codes of the NDM and codes for metacognitions that emerged from the data (see online Appendix E for details on the coding procedure). Through axial coding, we identified five decision making patterns that differed in the metacognitions they involved. Finally, theoretical coding allowed us to develop a process model of medical diagnosis decision augmentation with AI advice. The model reveals the underlying mechanisms that lead physicians to either follow AI advice or remain with their own initial frame.

Descriptive coding

We began by descriptively coding each sentence from the think-aloud protocols that included a new step in participants' reasoning. These codes were then arranged along a temporal process (see Table 4 for definitions), which helped us identify the metacognitive activities and the NDM process steps. We coded the decisions according to if participants changed or preserved their frame based on the last frame before they detected a problem. Based on the data analysis, three categories of metacognitions emerged. First, *self-monitoring* describes the subjective assessment of how well a cognitive task is, will, or has been performed by oneself (Ackerman and Thompson 2017). Two subcategories of *self-monitoring* are *feeling of rightness* and *personal decision process*. *Feeling of rightness* describes the conviction of the correctness of one's own frame based on the belief in personal capabilities to make a correct diagnostic decision, whereas *personal decision process* describes the calibration of one's confidence through the assessment of one's personal decision making process. Second, *system-monitoring* emerged as a subjective assessment of how

well a cognitive task is, will, or has been performed by the system. *System-monitoring* has two subcategories. *System capabilities* describes the conviction of the system's correctness based on the belief in the system's capabilities to make a correct diagnostic assessment, whereas *system inference process* describes the calibration of the perceived system accuracy through the assessment of the system's inference process. Finally, we identified three *metacognitive control* activities: *Frame elaboration* describes a phase of assessing the provided data after detecting a problem. *Inference control* refers to activities that control the personal decision making process—for example, ignoring information, purposefully considering the information in a specific order, or creating multiple contrasting combinations of pieces of information. *Asking for additional support* refers to decision makers asking for human expert advice or additional clinical information.

Table 4. Overview of descriptive codes

Concept and definition			**Sample code**
A priori defined codes based on the NDM framework			
Develop frame	Subtle complexes of data which lead to building the frame (Klein et al. 2005)	Accumulate cues about the patient case	"On the left, there are already inhomogeneous as well as hypodense areas" (Participant 35) (assessment of CT)
	The mental simulation of experiences and data into a mental model (Klein 2008). Usually the frame is built through pattern recognition	Develop preliminary hypothesis about diagnosis of COPD	"That's ... there's definitely COPD" (Participant 9)
Detect problem	The accumulation of discrepancies between the observation and desired states as well as the violation of expectations (Klein et al. 2005)	Detect that the provided information does not fit coherently	"So, I probably would have thought that was more of a COPD right now. Hm… but the tool tells me something else now" (Participant 22)
Preserve or change frame	As an outcome of the sensemaking activities, decision makers preserve or change the frame to a presumably better one (Klein et al. 2005)	Derive a diagnosis decision whether the patient has COPD or not. Coded as preserving or changing the frame based on the frame before the problem detection	"…so I'm saying NO COPD" (Participant 15)
Emerged codes for metacognition			
Self-monitoring	Subjective assessment of how well a cognitive task is, will, or has been performed by oneself (based on Ackerman and Thompson 2017)	1. Feeling of rightness: Consider subjective beliefs in own correctness (based upon Ackerman and Thompson 2017)	"Even though I'm not an expert..." (Participant 41)
		2. Personal decision process: Evaluate the correctness of the own decision making process based on the provided data	"But then I don't understand why the LAV should be 0%" (Participant 43)

System-monitoring	Subjective assessment of how well a cognitive task is, will, or has been performed by the system (based on Ackerman and Thompson 2017)	3. System capabilities: Consider subjective beliefs in CAID system accuracy	"It would be stupid not to believe the CAID since it has learned it" (Participant 31)
		4. System inference process: Evaluate the perceived accuracy of the AI based on the provided data	"And he [the AI] is still saying COPD now.… he is probably saying that now because of the -635 (mean lung density value)" (Participant 19)
Control	Initiating, terminating, or changing the allocation of effort to a cognitive task (based on Ackerman and Thompson 2017)	5. Frame elaboration: Consider data after problem detection. Often new pieces of information or previous sources	"But, in the introduction video, there was more severe emphysema in the lungs than in this picture" (Participant 40)
		6. Inference control: Activity to selectively include or exclude certain pieces of information	"I'm trying not to look at the AI advice this time so I don't get distracted." (Participant 2)
		7. Asking for additional support Asking for additional information or human expert advice	"So, if this were a clinical situation, I'd get some more expert advice." (Participant 20)

Pattern-identification

We used axial coding to identify patterns of metacognitive activities (see online Appendix E for details). For the pattern development, we combined evidence from the think-aloud protocols, survey data of each participant, post-experiment interviews, video recordings, and observations of the participants during the interaction with the CAID system. The patterns evolved through an iterative process of analyzing the qualitative and quantitative data. We incorporated data displays, tables with event frequencies, data matrices, and pattern matching (Miles and Huberman 1994). A total of five distinct patterns emerged that we elaborate in the results section.

Development of a process model

Next, we identified the theoretical mechanisms underlying the decision making process with AI advice (Strauss and Corbin 1990, Saldaña 2013). We used memos, transcripts, data matrices, and pattern visualizations (Miles and Huberman 1994 p.46). Through theory building, we developed a process model of medical diagnosis decision augmentation with AI advice. Our analysis revealed that decision makers must overcome two dynamic conflicts between *self-monitoring* and *system-monitoring*. We demonstrate that the way decision makers navigate through these conflicts influences not only the success of their interactions with AI advice but also their satisfaction and final confidence in the decision.

RESULTS

In the following section, we first categorize the decision outcomes in our study regarding the correctness of the final decision (accuracy rate). Subsequently, we describe patterns how physicians proceeded when AI advice confirmed or disconfirmed their assessments. Finally, we report the results of triangulating our findings with novice and experienced physicians.

Categorization of participants according to decision outcomes

Table 5 provides an overview of the participants' accuracy rates in diagnosing patient cases with correct and incorrect AI advice. In the control trial without AI advice, participants achieved an accuracy rate of 77%, which is comparable to the approximately 80% accuracy rate found in practice (Alberdi et al. 2004). Compared to the control trial, the accuracy rate for diagnoses supported by correct AI advice was marginally higher ($\chi2(1) = 3.05$, $p < 0.10$), whereas the accuracy rate was significantly lower for diagnoses with incorrect AI advice ($\chi2(1) = 9.31$, $p < 0.05$). The difference between diagnoses with correct versus incorrect AI advice was significant ($\chi2(1) = 19.84$, $p < 0.001$). In our experiment, the overall accuracy rate with CAID support was 72.09%. However, our CAID system provided 50% incorrect advice, which means that our system was significantly less accurate than CAID systems used in clinical practice. To position our results in relation to other studies and actual clinical practice, we extrapolated the overall accuracy rates based on our sample's accuracy rate in the context of correct and incorrect advice, respectively. For CAID systems providing correct advice 90% of the time, which corresponds with clinical practice (see, e.g., Jiang et al. 2017), the extrapolation indicates that the overall accuracy rate of our sample would rise to 86.88% (see online Appendix F). According to this extrapolation, novice physicians would clearly perform better with the support of a relatively accurate (90%) but imperfect CAID system than without any CAID system at all. As indicated in Table 5, participants experienced more disconfirmation from incorrect advice (Disconfirmation II) than from correct advice (Disconfirmation I). Eight participants felt confirmed by incorrect advice (Confirmation II). We found no evidence to suggest that performance in the incorrect AI advice trial was influenced by expertise or demographic variables: there was no significant difference between participants who decided correctly or incorrectly in their self-assessed expertise ($T(43) = -0.03$, p

= 0.98), their semester of study (T(43) = -0.65, p = 0.52), amount of clinical experience (T(43) = -0.65, p = 0.52) or demographic variables (see online Appendix G).

Table 5. Novice physicians' accuracy rates and occurrence of confirmation and disconfirmation

Experimental condition and number of participants		**Novice physicians with correct final diagnosis (accuracy rate)**	**Number of novice physicians with initial frame that was…**	
			Correct	**Incorrect**
AI advice	Control (No AI advice) (n=47)	76.60% (n=36)	NA	
	Correct (n=42)	90.48% (n=38)	Confirmation I 80.95% (n=34)	Disconfirmation I 19.05% (n=8)
	Incorrect (n=44)	54.55% (n=24)	Disconfirmation II 81.82% (n=36)	Confirmation II 18.18% (n=8)
Notes. Each participant (n=47) participated in all three trials and received one case of correct and one case of incorrect AI advice. Five participants in the correct advice trial and three participants in the incorrect advice trial reported not seeing the AI advice and were thus excluded from the analysis in this trial.				

Identified confirmation and disconfirmation patterns

The participants differed in their decision making and usage of metacognitions. Based on the analysis of our qualitative data, we identified five patterns that impacted accuracy rates, confidence, and satisfaction. In the unsupported trial, most participants remained with their first assessment. Thus, the decision accuracy was strongly dependent on the accuracy of the first, often intuitive assessment (see online Appendix H for more details).

Illustrating pattern 1: Confirmation

Confirmation is characterized by decision makers preserving their frame after a brief assessment of the provided data. We observed two different subpatterns of confirmation. First, participants evaluated provided data and then experienced a confirmation, which increased their confidence. Second, AI advice strongly influenced participants' assessments and frame building.

Pattern 1a. Data-based confirmation. When AI advice confirmed participants' assessments, confidence in their own assessment increased. As illustrated in the following example of Participant 11, the decision making process was straightforward and coherent and involved no jumping between features or cognitive activities. Participant 11 first assessed the provided information (Lines 1-9). Afterwards, the

participant built a frame determining that the patient suffers from COPD (Lines 10-12). Since the CAID confirmed the frame (Line 12), the physician preserved the original frame (Lines 13-14).

1	Ok, so the first thing I see is that the artificial intelligence brings up COPD as a diagnosis.	Accumulate cues
2	I am now considering the image	
3	and that already looks very...dark, so...yes.	
4	Taken together...quite large portions of rather dark tissue.	
5	Then I look at the values again,	
6	we have the mean lung density again,	
7	which is much too low.	
8	This time also the LAV is 46%,	
9	which is quite a lot,	
10	so, I would have already said it was COPD.	Build frame
11	And with the picture...I would also say it is COPD.	
12	And if the artificial intelligence also says COPD... then.	
13	Yes, the total volume is also increased.	Preserve frame
14	I would also say COPD.	

Physicians following this pattern considered all information and verified that all available data pointed in the same direction. This process increased their confidence and satisfaction with the decision. The accuracy was not determined by the AI advice but hinged on the correctness of the frame of the decision maker; the AI advice merely supported the physician's frame. However, some of our study participants indicated that they did not view a system as beneficial that just confirmed their own assessments.

Pattern 1b. AI-based confirmation. In contrast to Pattern 1a, AI advice also influenced how decision makers built their frame. Physicians following this pattern built their frame explicitly on the AI advice and did not sufficiently evaluate all the available information. For instance, Participant 31 received incorrect AI advice in the third trial while diagnosing a healthy patient. The CAID system incorrectly advised that the patient is suffering from COPD.

1	So, if I look again, the diagnosis [of the CAID] is COPD.	Develop AI-based frame
2	now I am looking at the lung CT...	
3	oh, or rather I didn't make (the diagnosis), the system made the decision	
4	Now I'm looking at the lung CT myself.	
5	So, at the base, it still looks relatively normal.	
6	If I go further up,	
7	you can already see that the tissue is partially hypodenic.	
8	But I find significantly less [hypodenic tissue] than in the last patient.	Feeling of rightness
9	So I personally think it would have been harder for me to diagnose	
10	if I hadn't had the support right now.	
…	…	
19	but I [would say it's COPD] and not as acute as the last patient.	Preserve frame

During the frame-development phase, participants following this pattern developed a weak frame that was strongly influenced by the AI advice (Lines 1-4). Participant 31, for example, did not use

background knowledge and did not recognize patterns in the data to develop a frame. In general, participants following this pattern perceived the AI advice to be very positive and supportive. For instance, Participant 14 described the influence of the CAID system in the follow-up interview: "(…) it influenced me a lot. At least it gave me a lot of certainty." Participants who followed this pattern did not critically reflect upon the suggestion of the AI advice. If the AI was correct, this resulted in an accurate final diagnosis. However, if the AI was incorrect, this pattern was highly problematic, as it yielded an incorrect diagnosis.

Identified disconfirmation patterns

If the AI did not confirm physicians' frame, they detected a problem and engaged in various metacognitions to resolve this discrepancy.

Illustrating pattern 2: Belief conformity

Facing disconfirmation, individuals who followed one of two *belief-conformity* patterns did not assess provided information but only evaluated their beliefs in their own reasoning capabilities against their beliefs in the system accuracy. There were two dimensions of this pattern: decision makers either believed strongly in their own capabilities and ignored disconfirming AI advice without further evaluation (*Pattern 2a: Ignoring the AI*) or they trusted the AI judgment much more than their own assessment and blindly followed the system advice (*Pattern 2b: Favoring the AI*).

Pattern 2a. Ignoring the AI. Physicians adhering to this pattern did not consider the AI any further in their reasoning if it disconfirmed their own assessments. For instance, after realizing that the AI disconfirmed his assessment, Participant 12 argued: "He [the AI] recommends NO COPD. But, I'd still say it is COPD." The participant did not provide any further reasoning or justification as to why the AI should be rejected. Whereas such behavior does empower physicians to overrule incorrect AI advice, it may be highly problematic in situations where the physician is wrong and the AI advice is correct. Few of the novice physicians in our study chose to ignore the AI when the advice was provided in parallel with all the other information. This pattern was more frequently found in the reversed-order sample and among experienced physicians (see section Triangulation below).

Pattern 2b. Favoring the AI. A more common pattern observed among novice physicians was *favoring the AI,* in which decision makers developed a correct frame but struggled with disconfirming AI advice because of a conflict between their beliefs in their own competence and their beliefs in the AI capabilities. For instance, Participant 20 encountered incorrect advice in the third trial when diagnosing a patient suffering from COPD. In this case, the AI advice incorrectly indicated "NO COPD". The participant first developed the correct frame that the patient suffered from COPD (Line 8) after assessing four different pieces of data. In Line 9, the participant then explicitly detected that the AI gave disconfirming advice. Participant 20 then engaged in metacognitive monitoring including both *system-monitoring* and *self-monitoring* (Lines 10-14), but did not further assess the data. She then changed her frame in favor of the AI advice.

1	Now I also see the...recommendation, which is relatively prominent.	Develop frame
…	…	
8	So, I probably would have thought that this was COPD right now.	
9	Hm...but the tool tells me something else now.	Detect problem
10	Of course, the question is whether I trust it or not.	Belief in system capabilities
11	Um, um...hm...Now I'm trying to judge how accurate the tool is	
12	and how accurate I am in my personal judgment.	Feeling of rightness
13	As a nonexpert for COPD, I would have trusted the tool, but I'm very unsure about it.	
14	So, if this were a clinical situation, I'd get some advice from an expert.	Ask for additional support
15	But right now, I have to say something, so I'm saying NO COPD.	Change frame

Through these activities, the participant aimed to resolve the conflict between her frame and the AI advice. The participant monitored her belief in the reliability of the tool and compared it to her "personal judgment." These activities resulted in being even more uncertain regarding her own judgment, although she had already developed a correct diagnosis in Line 8. In Line 14, her low confidence in the decision combined with high beliefs in the capabilities of the AI resulted in being unable to decide without additional human expert advice. Like Participant 20, half of the participants following this pattern indicated they would have liked to request additional human expert advice. However, in the context of this high-pressure set-up, Participant 20 eventually complied with the CAID suggestion. In her follow-up interview, Participant 20 stated that she perceived the CAID system to be more accurate than her own judgment because of the huge database it relied on. Furthermore, she admitted that she was very uncertain: "… um,

but the second time I was a bit confused, because I thought something else than what the tool said. And then I was just uncertain, who's right now, me or the tool?"

We observed a similar pattern in Participant 23 who followed incorrect AI advice, stating that she "would be not so sure now. But if in doubt, I would trust the AI." As indicated by Participant 16, the CAID system offered something to rely on and something on which a decision could be based: "…as soon as I have the feeling I am not completely informed, I'd rather rely on something I can stick to. So, it [the AI] has more influence [on my decision]." To summarize, individuals following this decision pattern tackled disconfirmation based on their beliefs in their own competence versus their beliefs in the system capabilities. These individuals did not try to validate the advice by using available data; rather, they relied on their beliefs that CAID systems are highly accurate. They were often insecure about their final assessments and many would have preferred to also consult a human expert. Neither pattern 2a nor pattern 2b involved participants exerting significant effort to deliberately verify which diagnosis was correct—their own assessments or the AI advice.

Illustrating pattern 3: Self-justification

Compared to the *belief-conformity* patterns, physicians following the *self-justification* pattern used additional control processes to elaborate on the provided data and to determine if their own assessment or the system was correct. However, these physicians only used the data to trace back their own reasoning processes without deeply considering the AI advice. Specifically, these physicians only reassessed the data that served as a foundation for their first frame and often tried to justify their assessments by reusing the same input data without considering new information. In this pattern, AI advice was often rejected without further elaboration on its accuracy and without aiming to understand the underlying reasoning. If one's frame is incorrect, this strategy leads to incorrect decisions. Furthermore, although many decision makers who applied this strategy did make correct decisions, they often reported less satisfaction and less confidence in their final diagnoses.

Illustrating pattern 4: System-justification

In contrast to *self-justification*, *system-justification* describes physicians who validate AI advice without comparing it to their own reasoning processes. *System-justification* can take two forms: (1) Physicians can validate system advice by searching for information supporting the AI advice, which leads them to ignore their own assessments and follow the AI advice (*supporting the AI*). (2) Physicians can come up with reasons that justify why they assume that the AI assessment is incorrect, which leads them to reject the advice (*explaining away*).

Pattern 4a: Supporting the AI. Decision makers who followed this pattern adjusted their own assessments to confirm the AI advice. During the decision making process, the AI advice served as a decision anchor that influenced how information was processed. For instance, when faced with disconfirming AI advice, Participant 11 did not discard the AI advice; instead, he developed an alternative explanation to mitigate the conflict between his frame and the AI advice. Participant 11 accumulated cues provided in the CT image (Lines 1-7) and then realized that the AI advice conflicted with his frame (Lines 8-9). However, instead of realizing that this conflict might have resulted from incorrect AI advice, the participant looked for an alternative explanation to resolve the problem (Line 10), namely that the patient had "emphysema" and not COPD. Emphysema can be considered a symptom of COPD, but it does not necessarily indicate that the patient suffers from COPD. Thus, by diagnosing "emphysema", the physician accounted for the fact that the patient was not healthy while also accommodating the AI advice of NO COPD. The participant decided on a milder, but incorrect diagnosis in order to follow the AI advice. Through this process, other information that was clearly suggestive of COPD was not adequately considered and, in practice, this could have resulted in withholding COPD treatment from the patient. From a metacognitive perspective, this participant considered only the possibility of his own failure, rather than the possibility of a CAID failure. He used metacognitive control based purely on *system-monitoring* to change his initial assessment so that it became commensurate with the AI advice.

…	[…various symptoms indicating COPD]	Detect problem
8	Ah, and artificial intelligence says it's NO COPD.	
9	And this LAV value...is still higher...than it should be in a healthy person.	Frame elaboration to support AI
10	Yeah... so it could just be emphysema.	
11	Um...but you see, so what was explained in the video, you can see that.	
12	So, these changes in the middle compared to the outside.	

13 Hm...so I think... that there is somehow a change… which does not look healthy.
14 It could be pulmonary emphysema now, which is relatively pronounced.
15 Whether this is COPD... yes, in theory it could be.
16 Well, let's say pulmonary emphysema rather than COPD. } Change frame

Pattern 4b: Explaining away. Physicians following this pattern addressed the disconfirmation by searching for indicators that could explain why the CAID may have provided incorrect advice in a specific patient case. For instance, Participant 17 was confronted with incorrect AI advice that conflicted with her own assessment. However, the decision conflict was quickly resolved when Participant 17 located one value in the table of predicted values that she deemed not to fit with the other values and could possibly explain why the CAID system gave incorrect advice (Line 14): "it [the AI] is probably saying this now because of the -635 (mean lung density value)." Similarly, Participant 22 resolved the disconfirmation by stating: "I would have thought now that the picture is simply incorrectly rendered so that the program does not recognize it as emphysema." However, in our experimental design, all values except for the AI advice were aligned to point toward the correct diagnosis and there were no rendering errors. Nevertheless, while finding a possible explanation helped physicians discard the AI advice as incorrect, this would be potentially harmful in cases of correct AI advice. To summarize, both patterns of system-justification relied on metacognitive control to resolve the disconfirmation but focused exclusively on justifying why the system developed its assessment. In *supporting the AI*, the system was thought to be correct, whereas in *explaining away,* the system advice was discarded as resulting from system malfunction. Both patterns lacked validation of the decision maker's own reasoning process. Thus, if the CAID system provided incorrect advice, *supporting the AI* resulted in an incorrect diagnosis decision. Similarly, if the CAID system provided correct advice, *explaining away* resulted in an incorrect diagnosis decision.

Illustrating pattern 5: Active consideration

The final pattern describes physicians who account for both their own assessment and the system advice by considering both perspectives equally. The pattern is characterized by an in-depth decision making process with multiple iterations. For instance, Participant 18 received incorrect AI advice in the third trial. The CAID incorrectly advised NO COPD.

1 Okay, we're looking at another CT scan. } Develop frame

…		
7	all this speaks for COPD...according to the table.	
…		
14	Ah the artificial intelligence says that it's NO COPD.	Detect problem
15	The CT picture looks pretty bad to me,	
16	it definitely doesn't look like a normal lung should,	
17	I don't agree with the artificial intelligence.	Personal decision process + inference control
18	I did my diagnostics without looking at it,	
19	consciously, to be honest.	
20	You can see that the...air pockets...are not as...blatant as in the picture before.	Frame elaboration
21	So, it is not so conspicuous now but, nevertheless, there are different densities…	
…		
27	Well, I'm just looking at the picture again and again from top to bottom.	Personal decision process
28	Of course, I got a little confused by the statement made by the artificial intelligence.	
29	But, it still does not look like a normal lung picture to me.	Preserve frame
30	and so, I'd say there's a pathological finding, more like COPD.	

The participant first analyzed the provided data in the table (Lines 1-6) and accumulated the necessary information to build a frame (Line 7) and concluded that the patient clearly suffered from COPD. Then, the participant became aware of the conflict between the AI advice and her frame because the CAID indicated NO COPD (Lines 13-17). This resulted in a phase of metacognitions, in which the participant described *inference control* and how she excluded the AI advice when building her frame (Lines 18-19). Through this process, she increased the validity of her reasoning in comparison to the AI advice. After elaborating the frame for a second time and probing all factors that could potentially coincide with the AI advice, she correctly rejected the advice. In total, three characteristics of Participant 18's decision making process explain why Participant 18 preserved her frame, even though it was contrary to the AI advice. First, the participant exerted *inference control* by combining different pieces of information and leaving out others (i.e., the AI advice) to probe her hypothesis from different angles (Lines 17-19). Second, the participant evaluated both her own reasoning and the system advice. She traced back her own reasoning and tried to understand why the system had made its assessment. Lastly, she engaged in neutral *frame elaboration*, equally considering both the AI advice and her original frame. Participants following this pattern developed differentiated attitudes toward the CAID system; they perceived it as supportive but also recognized the need to constantly monitor the system's decisions.

Similarities and differences in confidence and satisfaction across patterns

We used survey data to compare how the qualitative patterns differed in terms of their reported confidence and satisfaction (see online Appendix I). Regardless of the correctness of the advice, decision

makers reported lower confidence ($T(85) = -3.24$, $p < 0.001$) and satisfaction ($T(85) = -2.06$, $p < 0.05$) if their decisions were disconfirmed versus confirmed. Individuals following *data-based confirmation* were the most confident (M = 3.47, SD = 3.81) and among the most satisfied (M = 3.81, SD = 0.92) with their decisions and with their interactions with the CAID system. Although disconfirmation was associated with an average drop in confidence, this drop was not spread equally across all patterns. *Favoring the AI* resulted in the lowest confidence (M = 1.81, SD = 0.77) and satisfaction (M = 2.36, SD = 0.88). Furthermore, patterns primarily based either on *self-monitoring* or *system-monitoring* exclusively were associated with low confidence (M = 2.13, SD = 0.95 for *self-justification*; M = 2.80, SD = 0.87 for *supporting the AI*) and satisfaction (M = 3.00, SD = 1.21 for *self-justification;* M = 2.53, SD = 0.96 for *supporting the AI*). Interestingly, individuals who adhered to the *explaining away* pattern reported high confidence (M = 3.34, SD = 0.76) and satisfaction (M = 4.13, SD = 0.51) even though their assessments were primarily driven by *system-monitoring* alone and their assessments were not based on the data. Decisions based on a combination of both *system-monitoring* and *self-monitoring* (i.e., *active consideration*) were associated with high confidence (M = 3.30, SD = 0.80) and satisfaction (M = 3.58, SD = 0.70).

Triangulation

We triangulated our findings from two different perspectives. First, we corroborated the identified patterns in a follow-up experiment applying a reversed-order design. Second, we evaluated how experience influenced the identified patterns with a sample of experienced radiologists.

Reversed-order experiment

The accuracy rate of novice physicians in the reversed-order experiment did not differ from the main experiment in the control trial ($\chi^2(1) = 0.21$, $p = 0.65$), correct AI advice trial ($\chi^2(1)=0.43$, $p=0.51$), and incorrect AI advice trial ($\chi^2(1)=1.69$, $p=0.19$). Results of the second, reversed-order experiment generally corroborated the identified patterns of the main, parallel-order experiment (see Appendix B and online Appendix J for direct comparison). *AI-based confirmation* was less prominent in the reversed-order experiment than it was in the parallel-order experiment, whereas the *ignoring the AI* pattern occurred much more frequently in the reversed-order experiment across all groups of participants. Probably due to the

smaller sample size, we did not find any occurrence of *explaining away* in the reversed-order experiment. The triangulation through a reversed-order experiment allowed us to understand the nuances of the identified patterns and rule out concerns about the influence of our study design on the identified patterns.

Patterns with experienced radiologists

Overall, we found that the patterns we identified among novice physicians also occurred with our sample of experienced radiologists. However, differences emerged because of substantial experience of these radiologists in diagnosing CT images (see Table B.1 in the Appendix). First, we observed that experienced physicians based their frame more on pattern recognition than on the AI advice. Furthermore, we discovered that the experienced physicians considered more differentiated aspects and different alternative diagnoses when building their frame and deriving a final decision. We also observed that the experienced physicians more frequently chose to ignore disconfirming advice even if it was actually correct (Disconfirmation I). Thus, the accuracy rate in the correct AI advice trial did not differ significantly from that of the novice physicians ($G^2(1) = 0.44$, $p = 0.61$). In addition, if the disconfirming advice was incorrect (Disconfirmation II), experienced physicians used more *control* activities, i.e., *frame elaboration* and *inference control*, than novice physicians. However, we observed that the experienced physicians more frequently followed the *supporting the AI* pattern and developed a solution that was commensurate with both their own assessments and the AI advice. This process resulted, to our surprise, in a nonsignificant difference in the number of accurate results, as compared to the novice physician sample ($G^2(1) = 0.58$, $p = 0.53$). Finally, the accuracy rate of experienced radiologists in the unsupported trial was descriptively higher than that of novice physicians at 83.3% (10 out of 12), but was not statistically significant ($G^2(1) = 0.27$, $p = 0.72$).

ANALYTICAL SUMMARY: A PROCESS MODEL OF MEDICAL DIAGNOSIS DECISION AUGMENTATION WITH AI ADVICE

We analyzed similarities and differences among the five patterns (see Table 6) to develop a process model that depicts how physicians make diagnostic decisions augmented by AI advice and to understand what determines whether physicians accept or discard incorrect AI advice. The resulting model is displayed

in Figure 2. It proposes that the physicians' decision to either preserve or change their initial frame is based on (1) whether AI advice confirms or disconfirms their assessment *and* (2) on distinct configurations of metacognitions, which we depict as two metacognitive conflicts.

Interacting with confirming AI advice

In general, if physicians feel confirmed by AI advice, they preserve their frame without engaging in extensive metacognitive activities. The AI advice reduces uncertainty, increases physicians' final confidence in their diagnosis, and their satisfaction with the CAID system. A first crucial possibility for error occurs, however, if physicians develop their frame primarily based on received AI advice instead of extensive data analysis (*AI-based confirmation*). Physicians who develop an AI-based frame fully rely on the AI advice as an easy solution, in which they are also fully confident. This is particularly critical if the AI advice is incorrect as it easily persuades physicians to make incorrect diagnostic decisions. Alternatively, physicians can develop their frame through recognizing patterns in available data using their skills and intuition. Once they compare the AI advice with their independently developed frame, they perceive a confirmation only if the AI advice corresponds to the frame (*data-based confirmation*). Only if physicians engage in data-based frame building, they are able to detect possible incongruences between provided AI advice and other data.

Table 6. Differences and similarities of patterns regarding metacognitions

Pattern	**Description**	**Distinctive metacognitive activities**			**Experienced conflicts**	**Accuracy of decision**
		Self-monitoring	**System-monitoring**	**Control**		
1: Confirmation						
1a) Data-based confirmation	AI advice confirms frame based on pattern recognition	-	-	Cue accumulation	-	Typically accurate
1b) AI-based confirmation	Frame is built upon AI advice	-	-	-	-	Depends on CAID accuracy
Disconfirming AI advice						
2: Belief conformity						
2a) Ignoring the AI	AI advice is ignored without further evaluation	Strong feeling of rightness	Weak belief in system capabilities	-	Belief conflict	Depends on correctness of frame
2b) Favoring the AI	Strong beliefs in AI capabilities drive decision	Weak feeling of rightness	Strong belief in system capabilities	-	Belief conflict	Depends on CAID accuracy
3: Self-justification						

3) Self-justification	Data analyzed to support own assessment only	Focus on one's own decision making process	-	Frame elaboration	Validation conflict	Depends on correctness of frame
4: System-justification						
4a) Supporting the AI	Data analyzed to support AI assessment only	-	Develop understanding of AI inference	Frame elaboration	Validation conflict	Depends on CAID accuracy
4b) Explaining away	Data analyzed to find inference error of AI	-	Develop understanding of AI as error	Frame elaboration	Validation conflict	Depends on correctness of frame
5: Active consideration						
5) Active consideration	Neutral evaluation of all data	Focus on one's own decision making process	Develop understanding of AI inference	Frame elaboration + Inference control	Validation conflict	Typically accurate

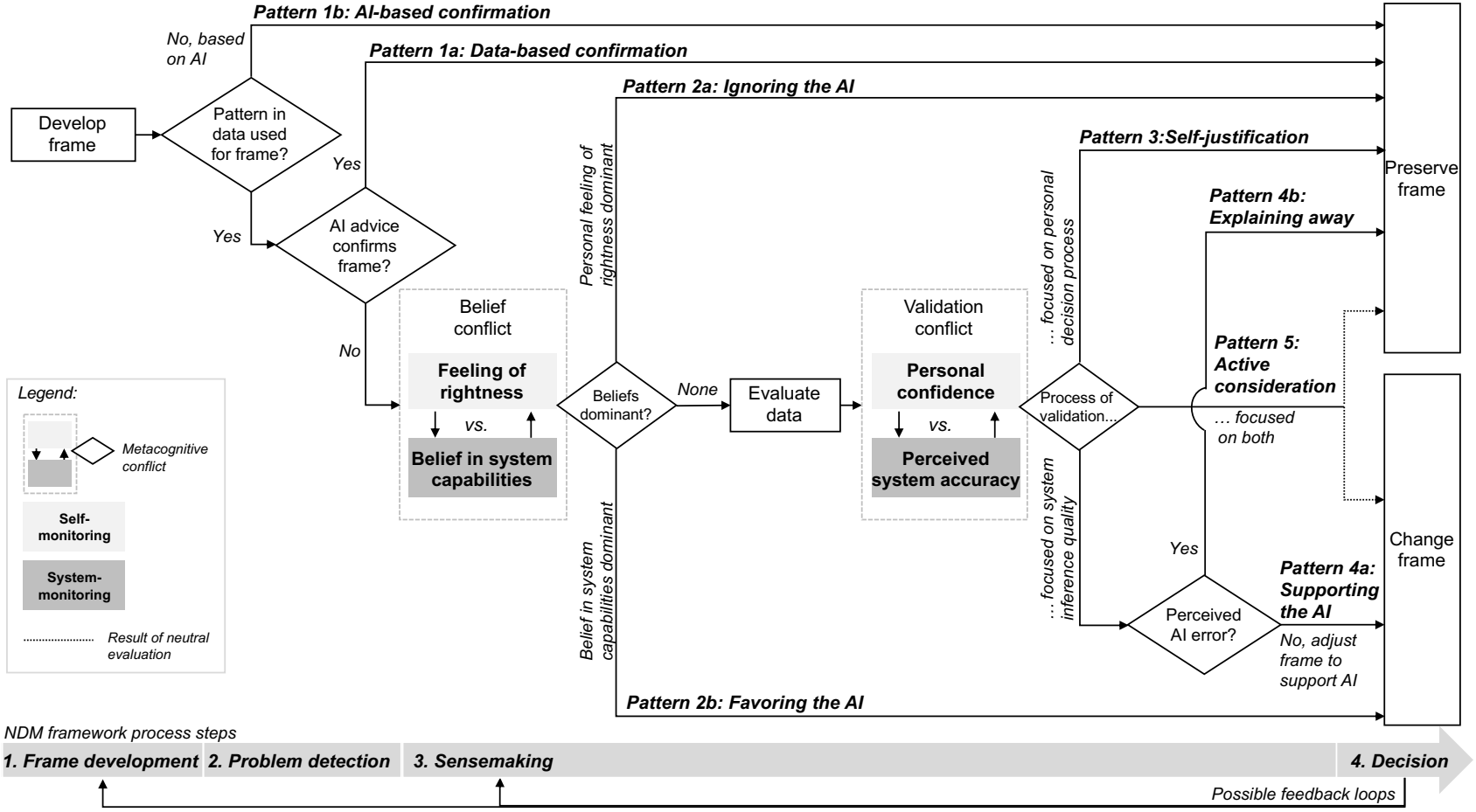

Figure 2. Process model of medical diagnosis decision augmentation with AI advice

Resolving two metacognitive conflicts in disconfirmation

By contrast, if the AI advice does not coincide with a physician's initial frame, they experience a disconfirmation of their assessment and, thus, they can either preserve or change their frame for concluding a diagnostic decision. The subsequent evaluation unfolds along a sequence of two metacognitive conflicts: the *belief* and *validation conflicts*. In the *belief conflict,* physicians base their decisions on rather stable beliefs about themselves and the system. More specifically, they compare their beliefs in the CAID system's capabilities to offer a correct diagnosis against their *feeling of rightness*, i.e., their belief in their own capabilities to make a correct diagnosis. Navigating this conflict, physicians immediately make a diagnosis decision if one of these beliefs is much stronger than the other. In this case, physicians do not consider any additional data but make a final diagnostic decision based on the dominant belief (*ignoring the AI* or *favoring the AI*). Especially in the *favoring the AI* pattern, physicians tend to remain mired in the *belief conflict* as their strong beliefs in the system's capabilities conflict with their frame. Therefore, they often lose confidence in their decision and try to consult additional expert sources.

If neither belief is dominant, physicians move into the *validation conflict,* in which they use additional data to validate their own frame and the AI advice. Aiming to resolve the *validation conflict*, decision makers continuously evaluate their personal confidence into their own frame against their perception of the system accuracy. Physicians preserve their frame if they are more confident in the correctness of their own frame in comparison with their perception of the AI advice accuracy. Otherwise they change their frame to correspond to the AI advice. Distinct metacognitive activities, namely *system-monitoring* and *self-monitoring,* determine the direction of this evaluation. As such, physicians who base the elaboration of their frame primarily on *self-monitoring* focus on their own decision process, trace back their own steps, and double-check whether the data fit their intermediary conclusions. These individuals often repeat their prior reasoning pathways and tend to overlook data elements that speak for the correctness of the AI advice (*self-justification*). By contrast, physicians who base the elaboration of their frame primarily on *system-monitoring* focus on the system inference quality and try to understand how the CAID system came to its conclusion. They search for data that provides proof for the accuracy or the inaccuracy of the AI advice. These individuals tend to either follow the AI advice

because they identify data that supports the AI advice (*supporting the AI*) or they dismiss the AI advice as a system failure because they identify potential errors in the data, which lowers their perception of the system's accuracy (*explaining away*). In either case, the physicians engage intensively with the AI advice but fail to compare the advice critically with their own assessment process. Only those physicians who reassess available data based on both *self-monitoring* and *system-monitoring* are able to conduct a neutral assessment (*active consideration*). Physicians engaging in active consideration compare the steps of their decision process with the data elements that support the AI advice and then refine their perceived confidence in the accuracy of their original frame and their perceived system accuracy. Overall, the degree to which physicians rely on *self-monitoring* and *system-monitoring* to navigate the two metacognitive conflicts determines if they preserve or change their frame in the face of disconfirming AI advice.

Accepting or discarding incorrect AI advice

Reacting appropriately to incorrect AI advice is highly challenging. In our study, we observed that the frequency of different patterns is related to the correctness of the AI advice and the expertise of the participants. Novice physicians more frequently fail to discard incorrect advice (Disconfirmation II) than experienced radiologists, whereas experienced radiologists more often ignore correct advice (Disconfirmation I). Our model outlines three decision pathways that lead physicians to inaccurate decisions given incorrect AI advice. First, decision makers can fail to detect a problem with incorrect AI advice since they base their own initial frames on the advice and thus perceive it as confirming (*AI-based confirmation*). We found this pattern to occur most frequently among medical students with little clinical experience. Second, physicians confronted with the *belief conflict* who have strong beliefs in the capabilities of a CAID system often unduly accept incorrect AI advice. Their strong beliefs in the system capabilities combined with doubt about their own capabilities reduce their decision confidence and leave them vulnerable to accepting incorrect advice (*favoring the AI*). Finally, individuals engaged in a *validation conflict* who reassess data primarily based on *system-monitoring* often fail to question the AI advice as they search for data that correspond to the advice (*supporting the AI*). In the context of incorrect AI advice, it is worth mentioning that individuals who *favor the AI* and *support the AI* initially

develop a correct frame but change it to correspond with the incorrect AI advice. Only *active consideration* results in successful interactions with both correct and incorrect AI advice.

DISCUSSION

Our study elaborates how physicians evaluate AI advice and how CAID systems influence their decision making process. Current research focuses primarily on how to leverage and improve AI capabilities to technologically support human decision makers in order to reduce human errors (e.g., Cheng et al. 2016, Kleinberg et al. 2018, Ahsen et al. 2019). Our study sheds light on a different but overlooked facet of decision augmentation with AI, namely the crucial role of human actors in compensating for technology errors. Decision makers such as physicians have the important role of mitigating the biases of AI-based systems and deciding whether AI advice should be transformed into concrete action. Although our study indicates that correct AI advice improves physicians' accuracy, our findings reveal three reasons that prevent physicians from reaping the full benefits of advice from CAID systems: First, the AI advice can influence physicians at the beginning of their decision making process. Thus, they may fail to conduct their own assessment of the data independently. Second, when AI advice disconfirms physicians' opinions, physicians are vulnerable to making incorrect diagnoses if they base their decision solely on their beliefs in their own capabilities or on their beliefs in system capabilities without further validation of available data. While more experienced physicians are more likely to ignore AI advice and disregard correct advice, novice physicians tend to lose confidence in their own assessments and are then prone to falling for incorrect advice. Overall, such belief-based decisions lead to lower confidence and less satisfaction with the system. Third, even if physicians exert metacognitive control and use data to validate received AI advice, the evaluation is often exclusively based on either *self-monitoring* or *system-monitoring*. Thus, physicians either attempt to reevaluate their personal decision process without sufficiently considering the AI advice or they try to find available data to support the system assessment without sufficiently considering that it may be incorrect. Yet, decision augmentation is most successful if decision makers draw on both *system-monitoring* and *self-monitoring* metacognitions to actively reconsider their own assessments as well as the accuracy of system advice. Our findings have a number of implications for research and practice.

First, from a theoretical perspective, we add to the literature by developing a process model of decision augmentation with system advice that accounts for the role of metacognitions. Our model integrates fragmented findings of prior research on system advice by identifying two metacognitive conflicts that are dynamically interrelated. As such, prior research suggested on the one hand that deliberate, in-depth reasoning is crucial when dealing with system advice (e.g., Heart et al. 2011) and on the other hand that decision makers fail to engage in such reasoning due to a number of heuristic biases towards systems as well as against systems (e.g., Adomavicius et al. 2013, Dietvorst et al. 2015). In doing so, prior work took a static perspective and regarded deliberate, systematic reasoning and superficial, heuristic reasoning in isolation from each other (Farrett et al. 2018). In contrast, our model shows that decision makers increasingly apply deliberate reasoning activities as they move from the *belief conflict* to the *validation conflict*. Beyond that, our metacognitive perspective helps to understand dynamic shifts between systematic and heuristic reasoning that can occur multiple times during the decision making process. For instance, even if decision makers deliberately evaluate system advice during the *validation conflict*, they can still be misled into a quick and superficial information search if they only engage in *system-justification*. Our model thereby accounts for the rising criticism on dual process theories as artificially dividing heuristic from systematic reasoning processes although decision makers often shift from one to the other dynamically (Evans and Stanovich 2013, Farrett et al. 2018). Moreover, we add to research on metacognitions by distinguishing *system-monitoring* from *self-monitoring* metacognitions. Whereas research on metacognitions has elaborated on the role of *self-monitoring* (Ackerman and Thompson 2017, Fiedler et al. 2019), we demonstrate that decision makers also monitor a system's performance. *Self-monitoring* and *system-monitoring* metacognitions can condition each other and their dynamic interplay influences decision outcomes. Our findings moreover suggest that system design can affect the prevalence of metacognitions and cognitive patterns. Specifically, physicians more often ignored disconfirming advice when the advice was presented after they had already made their own assessments. Future research should thus investigate how the use of metacognitions can be influenced by system design in a more nuanced way.

Second, our findings raise broader societal questions about the impact of AI on decision augmentation. Current discussions often revolve around the automation of tasks and if human experts

or AI-based systems should have primary agency in important decisions (Nissen and Segupta 2006, Demetis and Lee 2018, Rahwan et al. 2019). For AI-based systems in healthcare, this discussion was seemingly resolved when developers and physicians agreed that AI-based systems should only support the medical experts responsible for making final decisions (e.g., Cheng et al. 2016, Ahsen et al. 2019, Shen et al. 2019). Our findings demonstrate that this perspective neglects that human decision makers may lose major parts of their decision agency in the presence of AI-based systems without even noticing it. Specifically, our study suggests that if decision makers fail to engage in adequate metacognitive activities, they become cognitively unable to discard incorrect AI advice. Although final decision authority may rest with human decision makers, the actual decisions can be strongly impacted by AI advice. This can easily lead to a broader but less obvious substitution of human agency in crucial decision tasks than previously acknowledged. Thus, IS research should take cognitive factors in decisions with AI-based systems more openly into account. It should help to lead the societal discussion of whether this agency shift is desirable and how it can be regulated. Meanwhile, authorities that develop regulatory guidelines for the use of CAID systems in clinical practice need to recognize the cognitive challenges of physicians working with CAID systems. In particular, if CAID systems disconfirm physicians' assessments, it can be near impossible for physicians to reject AI advice if they are already involved in an intricate decision process. Future research on the use of AI-based systems needs to account for shifts in agency that users may not be aware of. Overall, our findings fuel the need to develop new theories, methods, and guidelines that account for this shift in agency (Nissen and Segupta 2006, Demetis and Lee 2018, Schuetz and Venkatesh 2020).

Third, our findings contribute to research on the interaction of physicians with clinical decision support systems. Without technological support, many medical errors occur because physicians often rely on their intuitive judgment and do not seek disconfirming information (Lee et al. 2013). Accurate CAID systems improve the overall accuracy of medical decisions and reduce medical errors (see Table 5). Although the overall decision making process may not differ between AI-based and rule-based systems, the distinct properties of CAID systems strongly influence *system-monitoring* metacognitions. As such, medical AI-based systems currently cause controversies regarding their actual performance versus expectations towards them (see Appendix A). Without prior experience with AI, many physicians

assume that AI-based systems work as a general AI that is able to cover a broad range of tasks autonomously. However, many current AI-based systems are narrow and specialized solutions that are trained to perform specific and relatively simple clinical tasks. In clinical practice, such expectations are therefore often violated and might contribute to an aversion to CAID systems (Dietvorst et al. 2015, Kohli and Tan 2016) and enforce decisions based on the *belief conflict*. Moreover, the current lack of transparency of AI-based systems creates difficulties in the *validation conflict* as monitoring the *perceived system accuracy* is more challenging compared to rule-based systems. With advances in explainable AI (Rai 2020), these difficulties will likely decrease as the reasoning behind AI advice becomes more comprehensible. In fact, there will likely be additional interactions between *system-monitoring* and *self-monitoring* as CAID systems become more explainable and powerful. As such, physicians may use provided explanations to learn from CAID systems and to gain new insights derived from the systems' pattern analysis. However, to benefit from these explanations, physicians need to further engage in *active consideration* as decisions based on the *belief conflict* would not result in increased knowledge and learning. Thus, future research needs to consider potential long-term feedback loops between different cognitive patterns and subsequent interactions with AI advice.

From a practical perspective, our findings can be utilized for training physicians in the use of CAID systems, especially in terms of dealing with disconfirming advice. Our results suggest that incorrect AI advice influences novice physicians in earlier stages of their decision making process than experienced physicians. Novices often fully base their diagnostic assessment on AI advice or follow the AI in the *belief conflict* whereas experienced physicians are more often influenced during the *validation conflict*. Thus, novice physicians must be trained in traditional data assessment skills and procedures to be able and willing to verify the accuracy of AI advice early on. For experienced physicians, it is important that they invest the additional effort of engaging in both *system-monitoring* and *self-monitoring* so that they do not unduly reject disconfirming advice but still spot occasional errors. In general, it is necessary for physicians to become aware of the different steps in their decision making process and to identify where mistakes are likely to happen. Otherwise, they may remain unaware that they are biased by provided advice. For example decision makers who based their initial assessments on AI advice did not consider that the advice might be incorrect at all. Moreover, we noticed that

physicians may fail to recognize the benefits of AI advice in confirmation if it "just" supports their own assessment. To prevent disconfirmed expectations, physicians should be made aware that confirmations through AI advice reduce ambiguity and uncertainty in a valuable way. To summarize, raising awareness of the cognitive mechanisms of the decision making process can help physicians to reap more benefits from diagnostic AI-based systems in the future.

Finally, our findings suggest that it is important to reconsider the role of radiologists who interact with CAID systems. Our findings do not suggest that replacing radiologists would actually lead to more accurate results. On the contrary, we argue that augmented intelligence is better than automation alone and we advocate for the consideration of both the human and technological aspects of medical diagnoses. Radiologists are increasingly becoming information specialists (Jha and Topol, 2016). In the future, radiologists may become orchestrators of multiple, distinct AI-based systems in different workflow steps, critical validators of system results, and translators between AI-based systems, clinicians, and patients. Although the speed and direction of AI technology development are unknown, it is necessary for radiologists to engage with AI-based systems in order to cope with increasing complexity, workload, and demographic changes in the clinical landscape.

Study limitations and future directions

Based on our choice of experimental design, our study has limitations that yield several paths for future research. First, our study participants were not allowed to gather additional information or seek human advice. However, we found that, especially in the *favoring the AI* pattern, the study participants desired human expert advice. Because of our study design, we do not know if decision makers wish to consult human experts in the case of disconfirmation because they explicitly distrust the AI technology (cf. Longoni et al. 2019) or if they would generally prefer another opinion. Moreover, interactions with human expert advice versus AI advice are likely to be very different. For instance, physicians can more easily dismiss disconfirming AI advice than human expert advice. Future research should therefore evaluate how physicians would integrate both sources of advice into their decision making process. Second, to examine the effects of AI advice on the decision making process, we included few details about our CAID system into the study design. Although there are AI technologies that are more transparent, black-box approaches based on deep learning such as ours are quite common,

particularly in radiology. Whereas our research provides insights into the latter area, future research should evaluate how variations such as providing explanations, explicit uncertainty and reliability rates, graphical information (e.g., coloring the relevant lung areas), or providing a choice as to whether to solicit system advice would lead to a different distribution of decision patterns. Third, compared to our experimental design, the decision making process in clinical practice is much more complex because decisions are often team-based and made under high time and resource pressure with numerous sources of clinical information (Burgess et al. 2010). Also, patients can suffer from multiple, interdependent conditions while symptoms and clinical data are often inconsistent. Therefore, future research should extend our insights and assess how these factors influence the process of medical diagnosis decision augmentation in clinical practice. Finally, although we propose that *active consideration* leads to the most elaborate decisions with AI advice, it is important to acknowledge that this pattern might not always be the most efficient solution in clinical practice. Specifically, decision making processes that reach the *validation conflict* need more time and effort than decisions made in earlier stages of the process. Considering the high workload of physicians, future research should reflect when and how physicians should optimally use *active consideration* to achieve the best results in clinical practice.

ACKNOWLEDGEMENTS

The authors thank Hemant Jain, Balaji Padmanabhan, Paul A. Pavlou, Raghu T. Santanam, special issue editors; Ben Shao, associate editor; and three anonymous reviewers for their constructive and helpful suggestions. We further thank Martin Pfannemueller, Felix M. Roth and Luis Oberste for their work in developing the CAID; Julia S. Beck, Marlene Buschlinger and Lukas Bossler for their support in the data collection and analysis; Likoebe M. Maruping, Thomas L. Huber and Mohammad H.R. Mehrizi for their valuable feedback and support in the paper development. We are grateful for the constructive feedback from participants at the special issue workshop, at the 6th Changing Nature of Work (CNoW) Workshop of the International Conference on Information Systems 2018, and at the Koenig Research Colloquium 2018. This research project is part of the Research Campus Mannheim Molecular Intervention Environment (M²OLIE) and funded by the German Federal Ministry of Education and Research (BMBF) within the Framework "Forschungscampus – public-private partnership for

Innovations". Lastly, we thank all participants in the experiments and especially the radiologists from the University Hospitals Mannheim and the Saarland University Medical Center, Germany.

REFERENCES

Ackerman RA, Thompson VA (2017) Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends Cogn. Sci.* 21(8):607–617.

Adomavicius G, Bockstedt JC, Curley SP, Zhang J (2013) Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Inform. Systems Res.* 24(4):956–975.

Ahsen ME, Ayvaci MUS, Raghunathan S (2019) When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Inform. Systems Res.* 30(1):97–116.

Alberdi E, Povyakalo A, Strigini L, Ayton P (2004) Effects of incorrect computer-aided detection (CAD) output on human decision making in mammography. *Academic Radiology* 11(8):909–918.

Arnold V, Clark N, Collier PA, Leech SA, Sutton S (2006) The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS Quart. 30*(1):79–97.

Arnott D, Pervan G (2014) A critical analysis of decision support systems research revisited: The rise of design science. *J. Inf. Technol.* 29(4):269–293.

Berg M (1997) *Rationalizing medical work: decision support techniques and medical practices* (MIT press).

Bonaccio S, Dalal RS (2006) Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organ. Behav. Hum. Decis. Process.* 101(2):127–151.

Burgess DJ (2010) Are providers more likely to contribute to healthcare disparities under high levels of cognitive load? How features of the healthcare setting may lead to biases in medical decision making. *Medical Decision Making* 30(2):246–257.

Burton JW, Stein MK, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *J. Behav. Dec. Making* 33(2):220-239.

Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM (2016) Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT Scans. *Nature Scientific Reports* 6(1):1–13.

Davern M, Shaft T, Te'eni D (2012) Cognition matters: Enduring questions in cognitive IS research. *J. Assoc. Inf. Syst.* 13(4):273–314.

Demetis D, Lee AS (2018) When humans using the IT artifact becomes IT using the human artifact. *J. Assoc. Inf. Syst.* 19(10):929–952.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych.: General* 144(1):114–126.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170.

Elkins AC, Dunbar NE, Adame B, Nunamaker JF (2013) Are users threatened by credibility assessment systems? *J. Manag. Inf. Syst.* 29(4):249–262.

Endsley MR (2017) From here to autonomy: Lessons learned from human-automation research. *Hum. Factors* 59(1):5–27.

Evans JStBT, Stanovich KE (2013) Dual-process theories of higher cognition: Advancing the debate. *Persp. Psych. Sc.* 8(3), 223–241

Fazal MI, Patel ME, Tye J, Gupta Y (2018) The past, present and future role of artificial intelligence in imaging. *Eur. J. Radiol.* 105:246–250.

Ferratt TW, Prasad J, Dunne EJ (2018) Fast and slow processes underlying theories of information technology use. *J. Assoc. Inf. Syst.* 19(1):1–22.

Fiedler K, Hofferbert J, Wöllert F (2018) Metacognitive myopia in hidden-profile tasks: The failure to control for repetition biases. *Front. Psychol.* 9(903):1–13.

Fiedler K, Hütter M, Schott M, Kutzner F (2019) Metacognitive myopia and the overutilization of misleading advice. *J. Behav. Decis. Mak.* 32(3):317–333.

Goddard K, Roudsari A, Wyatt J (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Informatics Assoc.* 19(1):121–127.

Hausmann D, Läge D (2008) Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgm. Decis. Mak.* 3(3):229–243.

Heart T, Zucker A, Parmet Y, Pliskin JS, Pliskin N (2011) Investigating physicians' compliance with

drug prescription notifications. *J. Assoc. Inf. Syst.* 12(3):235–254.

Jaspers MWM, Smeulers M, Vermeulen H, Peute LW (2011) Effects of clinical decision-support systems on practitioner performance and patient outcomes: A synthesis of high-quality systematic review findings. *J. Am. Med. Informatics Assoc.* 18(3):327–334.

Jha S, Topol EJ (2016) Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA - J. Am. Med. Assoc.* 316(22):2353–2354.

Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y (2017) Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* 2(4):230–243.

Kahneman D (2011) *Thinking, Fast and Slow*. New York, NY: Macmillan.

Kahneman D, Klein G (2009) Conditions for intuitive expertise: A failure to disagree. *Am. Psychol.* 64(6):515–526.

Kirkpatrick K (2016) Battling algorithmic bias. *Commun. ACM* 59(10):16–17.

Klein G (2008) Naturalistic decision making. *Hum. Factors* 50(3):456–460.

Klein G (2015) A naturalistic decision making perspective on studying intuitive decision making. *J. Appl. Res. Mem. Cogn.* 4(3):164–168.

Klein G, Pliske R, Crandall B, Woods DD (2005) Problem detection. *Cogn. Technol. Work.* 7(1):14–28.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *Q. J. Econ.* 133(1):237–293

Kohli R, Tan SSL (2016) Electronic health records: How can IS researchers contribute to transforming healthcare? *MIS Quart.* 40(3):553–573.

Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE (2013) Cognitive and system factors contributing to diagnostic errors in radiology. *Am. J. Roentgenol.* 201(3):611–617.

Li M, Tan CH, Wei KK, Wang K (2017) Sequentiality of product review information provision: An information foraging perspective. *MIS Quart.* 41(3):867–892.

Liberati EG, Ruggiero F, Galuppo L, Gorli M , González-Lorenzo M, Maraldi M, Ruggieri P, et al. (2017) What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implementation Science* 12(113):1–13

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human

judgment. *Organ. Behav. Hum. Decis. Process.* 151:90–103.

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *J. Consum. Res.* 46(4):629–650.

Mayo RC, Leung J (2018) Artificial intelligence and deep learning – Radiology's next frontier? *Clininical Imaging* 49:87–88.

Mayo RC, Kent D, Sen LC, Kapoor M, Leung JWT, Watanabe AT (2019) Reduction of false-positive markings on mammograms: A retrospective comparison study using an artificial intelligence-based CAD. *J. Dig. Imag.* 32:618–624.

Miles MB, Huberman AM (1994) *Qualitative Data Analysis: An Expanded Sourcebook* (Sage, Thousand Oaks).

Nissen ME, Segupta K (2006) Incorporating software agents into supply chains: Experimental investigation with a procurement task. *MIS Quart.* 30(1):145–166.

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. *Hum. Factors* 52(3):381–410.

Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon JF, Breazeal C, Crandall JW, et al. (2019) Machine behaviour. *Nature* 568(7753):477–486.

Rai A (2020) Explainable AI: from black box to glass box. *J. Acad. Mark. Sc.* 48(1):137–141.

Russell SJ, Norvig P (2010) *Artificial Intelligence: A Modern Approach* (Pearson Education Limited, Malaysia).

Saldaña J (2013) *The coding manual for qualitative researchers* (Sage).

Schuetz S, Venkatesh V (2020) The Rise of Human Machines: How Cognitive Computing Systems Challenge Assumptions of User-System Interaction. *J. Assoc. Inform. Systems* 21(2):460–482.

Schultze T, Mojzisch A, Schulz-Hardt S (2017) On the inability to ignore useless advice: A case for anchoring in the judge-advisor-system. *Exp. Psychol.* 64(3):170–183.

Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, He Z, et al. (2019) Artificial intelligence versus clinician in disease diagnosis: Systematic review (Preprint). *JMIR Med. Informatics* 7(3):e10010.

van Someren MW, Barnard YF, Sandberg JAC (1994) *The think aloud method: A practical guide to modelling cognitive processes* (Academic Press, London).

Strauss A, Corbin J (1990) *Basics of Qualitative Research: Grounded theory procedures and techniques,* 2nd ed. (Sage).

Tan CH, Teo HH, Benbasat I (2010) Assessing screening and evaluation decision support systems: A resource-matching approach. *Inform. Systems Res.* 21(2):207–412.

Todd P, Benbasat I (1991) An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies. *Inform. Systems Res.* 2(2):87–115.

Tsai TL, Fridsma DB, Gatti G (2003) Computer decision support as a source of interpretation error: The case of electrocardiograms. *J. Am. Med. Informatics Assoc.* 10(5):478–483.

Tversky A, Kahneman D (1974) Judgment under ucnertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Wang X, Du X (2018) Why does advice discounting occur? The combined roles of confidence and trust. *Front. Psychol.* 9(2381):1–13.

World Health Organization (2018) World Health Statistics. *World Heal. Organ.* Retrieved (July 10, 2020), https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Xiao B, Benbasat I (2011) Product-related deception in e-commerce: a theoretical perspective. *MIS Quart.* 35(1):169-196.

Xiao B, Benbasat I (2015) Designing warning messages for detecting biased online product recommendations: An empirical investigation. *Inform. Systems Res.* 26(4):793–811.

APPENDIX

Appendix A – Conceptual similarities and differences between AI-based (CAID) and rule-based computer-aided diagnosis systems

Table A.1. Conceptual similarities and differences between AI-based (CAID) and rule-based computer-aided diagnosis systems

	AI-based computer-aided diagnosis systems for radiology[1]	**(Traditional) rule-based computer-aided diagnosis systems for radiology[2]**
Aim of the algorithm	- Provide second opinion/advice - Automating workflows	- Provide support in process and decisions - Standardizing workflows
Knowledge base	- Inference from large data sets relatively independent from human expertise - Algorithms can develop new inferences	- Externalized knowledge and experience of human experts - Algorithms are static
Algorithm properties	- Learning: rules and conditions are not predefined, but emerge from data - Algorithms are often stabilized for usage in clinical practice; learning is restricted to model training - Inference logic neither transparent nor explainable (esp. deep neural networks) - Heterogeneous sources of information can be integrated more easily	- Rule-based: rules and conditions are predefined and must be manually expanded, often applied to standardized procedures and criteria - Inference logic is transparent and can be traced and explained - Integration of heterogeneous sources very challenging as it requires manual adjustment of rules
Algorithm performance	- Errors are not predicable as they result from biases in training data - Higher accuracy and efficiency	- Errors follow static patterns - Often high numbers of false positives
Role of the human	- Design, train, and implement the algorithm in clinical practice - Select meaningful parameters and context in the algorithm development - Provide classified data and ensure the quality of the data - Usage in clinical practice: interpret and validate the output of the system and use it as a second opinion	- Design and implement the algorithm in clinical practice - Specify the problem and extract and categorize human knowledge to develop the algorithm - Define conditions and rules for the algorithm (incl. boundaries and exceptions) - Update the knowledge base on a regular basis - Usage in clinical practice: follow a standardized procedure with multiple interaction points (e.g., enter input data); validate the process and output
Challenges	- Data basis: the "ground truth" is often determined by physicians, which involves high manual effort, errors, and biases - Data availability: in practice, it is often difficult to integrate new data because of inconsistent data quality - Specialized AI vs general AI: high expectations of AI performance often stem from a belief in the accuracy of general AI; however, current AI solutions often perform narrow tasks and are highly specialized - Interpretation of outcome: low transparency and explainability; psychological challenges of evaluating advice when it conflicts with one's own assessment	- Inflexibility of the algorithms: many boundaries in clinical decision making - Integration into the workflow: multiple steps of interaction are necessary and can change existent workflows - Interaction with the physician: motivating the physician to use the support is challenging and often unsuccessful - Electronic health records as data basis: challenges of data standardization and integration

Notes. This overview was developed based on expert interviews and literature synthesis of Berg 1997, Cheng et al. 2016, Jiang et al. 2017, Fazal et al. 2018, Ahsen et al. 2019, Mayo et al. 2019, Rahwan et al. 2019 and Shen et al. 2019

1. *We only focus on current systems in radiology, which are digital systems with no physical representation (in contrast to, e.g., robots). These systems mainly have two AI properties: machine learning (often with deep neural networks) and image recognition. This means that many AI capabilities (for example, natural language processing and speech recognition) are not considered in this context.*
2. *The conceptual line between AI-based systems and rule-based systems is often blurred as there are different understandings of intelligence. Thus, this differentiation is conceptual in nature.*

Appendix B – Occurrences of patterns based on level of experience and experimental design

The following Table B.1 and Table J.1 in the online Appendix compare the pattern occurrences based on different degrees of expertise. For these analyses, we split the sample of novice physicians into novice physicians without clinical experience and novice physicians with clinical experience. Through clinical training, novice physicians gain practical experience in diagnosing and thus, in comparison to physicians with no clinical experience, their reasoning process will be closer to that of experienced radiologists. All groups (novice physicians with and without clinical experience as well as experienced radiologists) made more inaccurate diagnosis decisions when provided with incorrect AI advice than without advice. Except *explaining away,* all patterns from the original parallel-order experiment also occurred in the reversed-order experiment. With higher degrees of experience, physicians used more data-based patterns. In addition, novice physicians with clinical experience reported more satisfaction and confidence if they followed the self-justification pattern, in comparison to novice physicians without clinical experience. In our experiment, experienced radiologists did not show *AI-based confirmation* patterns. Thus, neither the sample characteristics nor the experimental order we chose for the original parallel-order experiment systematically influenced the appearance of the identified patterns, but they did influence their frequency. More details can be found in online Appendix J.1.

Table B.1. Comparison of patterns based on expertise

		Novice physicians (without clinical experience) of main experiment	**Novice physicians (with clinical experience) of main experiment**	**Experienced radiologists**
Mean accuracy rates	Control	80.78%	71.43%	83.33%
	Correct AI advice	86.96%	94.74%	83.33%
	Incorrect AI advice	45.83%	65.00%	66.67%
Major differences in sensemaking	Confirmation patterns	Data (1a) and AI-based confirmation (1b)	Data (1a) and AI-based confirmation (1b)	Data-based confirmation (1a)
	Disconfirmation patterns	- Favoring the AI (2b) - Self-justification (3) - Very few cases of system-justification (4a and 4b) and active consideration (5)	- Favoring the AI (2b) - Explaining away (4a) and Confirming AI (4b) - Active consideration (5)	- Ignoring the AI (2a) - Supporting the AI (4b) - Active consideration (5)
	Cause for inaccurate decision	- AI-based confirmation for incorrect advice (1a) - Favoring the AI (2b)	- Favoring the AI (2b) - Supporting the AI (4b)	- Ignoring of correct AI advice (2a) - Supporting the AI (4b)