Bus Inf Syst Eng 64(3):293–309 (2022)
https://doi.org/10.1007/s12599-022-00750-2

RESEARCH PAPER

Radiologists' Usage of Diagnostic AI Systems

The Role of Diagnostic Self-Efficacy for Sensemaking from Confirmation and Disconfirmation

Ekaterina Jussupow · Kai Spohrer · Armin Heinzl

Received: 28 June 2021 / Accepted: 28 February 2022 / Published online: 22 April 2022

Abstract While diagnostic AI systems are implemented in medical practice, it is still unclear how physicians embed them in diagnostic decision making. This study examines how radiologists come to use diagnostic AI systems in different ways and what role AI assessments play in this process if they confirm or disconfirm radiologists' own judgment. The study draws on rich qualitative data from a revelatory case study of an AI system for stroke diagnosis at a University Hospital to elaborate how three sensemaking processes revolve around confirming and disconfirming AI assessments. Through context-specific sensedemanding, sensegiving, and sensebreaking, radiologists develop distinct usage patterns of AI systems. The study reveals that diagnostic self-efficacy influences which of the three sensemaking processes radiologists engage in. In deriving six propositions, the account of sensemaking and usage of diagnostic AI systems in medical practice paves the way for future research.

Keywords Artificial intelligence · Decision making · Medicine · Expert · Usage

Accepted after one revision by the editors of the special issue.

E. Jussupow (✉) · A. Heinzl
Universität Mannheim, Mannheim, Germany
e-mail: jussupow@uni-mannheim.de

A. Heinzl
e-mail: heinzl@uni-mannheim.de

K. Spohrer
Frankfurt School of Finance and Management,
Frankfurt am Main, Germany
e-mail: k.spohrer@fs.de

1 Introduction

Information systems based on artificial intelligence (AI) are changing the work of knowledge professionals by performing knowledge tasks that were previously considered too complex and too unstructured for effective information systems support (Faraj et al. 2018). A prime example is medical decision making in radiology, where radiologists examine images from computed tomography, magnetic resonance imaging, and X-ray to diagnose whether and to which extent patients suffer from diseases. To be able to conduct such important diagnoses correctly and reliably, radiologists must undergo years of training and education, including training in general medicine and an extensive period of specialization (Pratt et al. 2006). In this context, AI systems have been introduced that develop diagnostic assessments and perform on par with medical experts when it comes to diagnosing diseases such as stroke from radiological images (Hosny et al. 2018; Shen et al. 2019). These AI systems are currently implemented to assess patient cases in parallel to the evaluation by radiologists. It is then subject to the radiologists' judgment whether and how to include the diagnostic assessment from the AI system in a final, combined diagnosis. As AI systems increasingly support critical radiological decision tasks where life and health of human patients are at stake, it becomes even more important that radiologists use such powerful AI systems to support their diagnostic decision making.

However, several reasons suggest that it cannot be taken for granted that radiologists fully benefit from AI systems in their decision making. First, the medical profession has a history of rejecting support of new information systems, with physicians often playing a gatekeeper role (Lapointe and Rivard 2005). Physicians' decision to use or to reject a

novel information system has massive consequences for other clinical staff and their system usage (Romanow et al. 2018). Second, many AI systems in radiology rely on deep learning algorithms to classify imaging data (Jiang et al. 2017). Although they can constitute powerful tools, these systems are typically black boxes with no or very limited information for radiologists to understand how the system has come to its assessment (Fazal et al. 2018). Consequently, a radiologist's decision how to use the system and whether to include its assessment in a final diagnosis depends largely on the result of the AI system's classification and is driven by cognitive processes we are only beginning to understand (Fügener et al. 2021; Jussupow et al. 2021). Third, there is increasing evidence that knowledge professionals have a difficult time effectively integrating advice into complex decisions that is provided by an information system due to an aversion toward such systems (Dietvorst et al. 2015). Understanding these cognitive mechanisms and their ramifications constitutes an emergent field of vibrant research activity (Burton et al. 2020; Jussupow et al. 2020). Overall, this suggests that it is important for medical practice to further examine and understand how radiologists use AI systems to enhance their decision making.

From a theoretical point of view, the processes how physicians come to use information systems is particularly hard to understand because physicians are known as a user group that primarily decides about future system use based on implicit decisions during current use (Burton-Jones and Volkoff 2017). Whereas other knowledge professionals often possess some slack resources and time to explicitly reflect on the possible use of new information systems, Burton-Jones and Volkoff (2017) emphasize that physicians primarily reflect in action by evaluating their current interactions with a system during task performance, implicitly making sense of these interactions to shape their future system use. This is particularly interesting in the context of diagnostic AI systems in radiology. In evaluating diagnostic AI advice radiologists gain one primary piece of information, namely whether the system's assessment of a patient case confirms or disconfirms their own judgement (Jussupow et al. 2021). Consequently, understanding how radiologists come to use an AI system in clinical practice hinges on a better understanding of how they handle confirmation and disconfirmation.

However, the burgeoning research on the impact of AI systems on knowledge work (Faraj et al. 2018; Kellogg et al. 2020) and decision making (Burton et al. 2020; Fügener et al. 2021; Jussupow et al. 2021) has not yet investigated how radiologists come to use AI systems in clinical practice. Although a number of experimental studies suggest that individuals' acceptance of AI advice in isolated decision making tasks is subject to intricate cognitive processes and that confirming AI advice triggers different cognitions than disconfirming advice (Dietvorst et al. 2015; Jussupow et al. 2021), it has so far remained unclear how these findings translate to actual usage of AI systems in medical practice beyond isolated decision making tasks. Against this backdrop, we address the following research questions:

RQ1: How do radiologists differ in their usage patterns of AI systems for diagnostic decision making?
RQ2: What role do AI interactions that confirm and disconfirm the radiologists' own judgment play in forming usage patterns?

To answer these questions, we present the results of an exploratory case study at a German University Hospital in which an AI system is implemented to diagnose stroke. Drawing on observations and interviews, we take a sensemaking lens and show how radiologists make sense of confirmation and disconfirmation by the AI system, using the system in different ways. We identify three sensemaking processes that constitute the basis of distinct usage patterns. We outline how all three of them critically depend on radiologists' diagnostic self-efficacy and can have feedback effects on the same. We contribute to extant knowledge by presenting a preliminary model of physicians' sensemaking and usage of diagnostic AI systems and by deriving six propositions that we hope will spur further research.

2 Conceptual Foundations

2.1 Impact of AI Systems on Medical Work

AI systems are considered as systems which resemble human abilities in reasoning, generalizing, or learning from experience (Russell 2019; Russell and Norvig 2010). Those system shift the relationship between users and technology (Baird and Maruping 2021) as they strongly influence how decisions are made and challenge the supremacy of human expertise (Faraj et al. 2018). In medical work, technological advances in general have changed the historical definition of good, rational medical decisions from an expert-driven intuition to a more data-based decision making (Berg 1997). With rapid implementation of advanced AI systems, the overall quantification of knowledge work increases (Faraj et al. 2018; Kellogg et al. 2020). Thus, instead of being the sole decision maker with unquestioned decision autonomy, the quality of diagnostic decisions of medical experts is now sometimes compared with the accuracy rate of AI systems. Meanwhile, patients have become increasingly aware of AI systems and are judging medical professionals depending on how they evaluate AI

system advice (Arkes et al. 2007; Longoni et al. 2019; Shaffer et al. 2013).

Particularly in radiology, medical specialists are conducting tasks that are already highly quantified, as they work with computed tomography scans, magnetic resonance images, and quantified indicators derived from those images (Hosny et al. 2018). Thus, in this domain, AI systems are already performing with high accuracy and efficiency in segmenting and classifying images (Hosny et al. 2018). This increasing pressure from AI systems changes how radiologists work and forces them to redefine their professional role (Tang et al. 2018). In fact, Geoffrey Hinton caused uproar in the radiology community as early as 2016 by stating: "We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists" (Hinton 2016). However, while hospitals face resource shortages in terms of personnel and time, the increasing personalization and quantification of diagnostic and treatment decisions requires more human and technology resources for each individual patient. Consequently, most radiologists agree that it is necessary to adopt AI systems in order to cope with the increasing workload, case complexity, and required diagnostic accuracy, but also believe that these systems may profoundly change their work in not yet determined ways (Miller and Brown 2018; Tang et al. 2018).

2.2 Usage and Impact of AI Systems in Medical Diagnostic Decision Making

Although increasingly popular, research on the impact of AI systems on knowledge work (Faraj et al. 2018; Kellogg et al. 2020; Sturm et al. 2021) and decision making (Burton et al. 2020; Jussupow et al. 2021; Fügener et al. 2021) has not yet investigated how radiologists come to use AI systems in clinical practice. Instead, prior work has mostly focused on isolated decision tasks, mostly in laboratory experiments. Nonetheless, this line of work constitutes an important basis for understanding AI system usage in radiological practice because physicians are known to form views and usage patterns of information systems primarily in action, specifically, during task performance (Burton-Jones and Volkoff 2017). In fact, usage patterns can be seen as an emergent result of physicians' reflections on an information system's affordances and its performance during task work (Burton-Jones and Volkoff 2017). Thus, understanding radiologists' reflections on an AI system is an important step towards understanding their AI system usage.

Two results from prior work on the impact of AI systems on knowledge work and decision making appear noteworthy. First, prior work suggests that successful use of AI advice in decision making tasks is influenced by how confident decision makers are about their ability to make a correct decision without a supporting information system. For example, Fügener et al. (2021) show that the decision to delegate a classification task to an AI system is driven by users' perceived ability to perform the task. Specifically, users only delegate a task to the AI system if they are not confident that they can perform the task well without support. However, the study also indicates that this perception is often biased, as humans do not accurately perceive their own confidence but tend to overestimate it. Furthermore, Dietvorst and colleagues (Dietvorst and Bharti 2020; Dietvorst et al. 2015) show that aversion toward algorithmic advice is driven by users perceived relative confidence into their own abilities versus the perceived accuracy of the algorithm. In particular, after seeing that an algorithm has erred, users are more confident into their own abilities to make a correct decision than in the algorithm. Also, Jussupow et al. (2021) indicate that monitoring one's own abilities influences the cognitive evaluation of AI advice and the decision to follow or reject disconfirming AI advice. Radiologists who are not confident during a diagnostic task tend to follow disconfirming AI advice more frequently. Although recent findings indicate that the learning ability of AI systems may partly compensate for negative impacts of prior system errors (Berger et al. 2021), imperfect system assessments often result in a loss of trust into the system and its accuracy (Dietvorst and Bharti 2020; Dietvorst et al. 2015). However, the dynamics how such beliefs play into the formation of long-lasting usage behaviors beyond isolated decision tasks are not yet clear.

In our study, we differentiate two forms of confidence. On a decision task level, there is *diagnostic confidence*, defined as a decision maker's perception how certain they are to make the right decision drawing on their own analysis of the specific decision task (e.g., Jussupow et al. 2021). Across multiple decision tasks, there is self-efficacy, defined as an individual's belief in their capacity to execute behaviors necessary to produce specific performance attainments (Bandura 1997). More specifically, we refer to *diagnostic self-efficacy* as physicians' belief in their capabilities to make correct diagnostic decisions.

A second relevant result from prior work is that different cognitive processes are triggered if AI advice confirms decision makers' assessment than if AI advice disconfirms their assessment (Jussupow et al. 2021). In the domain of radiology, Jussupow et al. (2021) demonstrate that experiencing disconfirmation by an AI system can, but does not necessarily, trigger cognitive activities that help decision makers determine whether their own judgment is accurate. In fact, radiologists can often be persuaded into incorrect decisions by AI systems although they would decide

differently without the influence of a disconfirming AI (Jussupow et al. 2021). To successfully navigate disconfirming AI advice, decision makers need to utilize more elaborate reasoning than for confirming AI advice in order to detect reasons for the divergent assessments and act accordingly (Kahneman and Klein 2009; Klein et al. 2007). Thus, if we want to understand radiologists' emergent usage of AI systems, we should consider potential effects of their confidence and to how they make sense of AI assessments that confirm or disconfirm their own evaluations of a patient case.

2.3 Sensemaking as a Theoretical Lens

We analyze our research problem through a lens of organizational sensemaking. Although sensemaking only emerged from our data analysis as a fitting theoretical framework, we elaborate on it at this point to facilitate the understanding of our findings. Sensemaking refers to constructing and reconstructing meaning, interpreting, and updating cognitive frameworks (Gioia and Chittipeddi 1991; Jenkin et al. 2019). Sensemaking is triggered by events that cause uncertainty for individuals, including changes in the organizational environment (Weick et al. 2005), threats to one's identity (Petriglieri 2011), and the introduction of information systems that may change one's work (Tan et al. 2020). From a sensemaking perspective, "using a technology is a cognitive process by which users construct meaning of the technology, which affects their subsequent interactions with it" (Hsieh et al. 2011, p. 2018). We are particularly interested in how radiologists make sense of their interactions with AI systems and the subsequent usage patterns that emerge from this sensemaking process.

There are three distinctive sensemaking activities that have been found helpful in explaining individuals' actions in and contributions to enterprise system implementations (Tan et al. 2020) and distributed work with information systems (Vlaar et al. 2008): sensedemanding, sensegiving, and sensebreaking. *Sensedemanding* refers to individuals' activity to acquire and process information to ameliorate uncertainty and equivocality (Gioia and Chittipeddi 1991). *Sensegiving* refers to individuals' activities that attempt to influence others' sensemaking activities toward a preferred interpretation of organizational reality (Vlaar et al. 2008). *Sensebreaking* refers to individuals' attempt to break and destroy meaning in order to induce new ways of thinking (Tan et al. 2020).

3 Method

In order to address our research questions, we conducted an exploratory single case study that can be considered as revelatory (Yin 2009). Exploration based on revelatory cases is particularly suitable for phenomena with little extant research (Sarker et al. 2012), which is the case for questions of how radiologists use AI systems for complex diagnostic tasks. Our case revolves around a productive AI system for stroke diagnosis in a radiology department of a German University Hospital. The case context was particularly suitable for understanding radiologists' AI usage because the AI system performed a diagnostic assessment in parallel to the radiologists' and had been established in the clinical routine at the hospital for two years. Thus, the local radiologists had had time to accustom to the system, reflect on it in action and routinize their usage practices.

We gained broad data access that allowed us to develop detailed insights into the usage of the system. We observed how radiologists interacted with the system and conducted interviews with them. Given our research questions, we wanted to elicit radiologists' account of their interactions with the system, their reactions to confirmation and disconfirmation, and the emergence of usage patterns. In line with quality criteria for exploratory case studies (Sarker et al. 2018), we gained rich and authentic accounts of the radiologists' situation and reasoning, drawing on their opinions, pleas, and confessions, thus reconciling the "polyphonic narrative" that became visible (Sarker et al. 2018). For data analysis, we borrowed elements of the grounded theory methodology (Saldaña 2013; Wiesche et al. 2017) that helped us make sense of what we observed and heard. In so doing, we conducted an exploratory study that helps uncover and understand how radiologists make sense of and use AI systems. Theorizing based on emergent views in the data and based on some pre-existing conceptions (i.e., the existence of confirmation and disconfirmation), our approach can best be described as abductive, acknowledging that our own pre-conceptions and thinking did form a major part of our analysis (Sarker et al. 2018, p. 759). Specifically, we took existing studies on the topic into account and even worked with a limited set of predefined concepts that helped us approach and structure the phenomenon while still working primarily based on the qualitative data. Finally, we derive a set of propositions from our qualitative investigation, which is in line with prior exploratory research using a sensemaking lens (Vlaar et al. 2008) and with recent socio-technical research that aims to understand individual interpretations and reactions to technology in a healthcare context (Califf et al. 2020).

3.1 Case Setting

The University Hospital was one of the first hospitals in Germany that introduced AI systems into clinical routine. The AI system was created by a company founded in 2010 and uses a supervised machine learning approach to classify the severity of stroke by automatically generating the Alberta Stroke Program Early Computed Tomography (ASPECT) score from computed tomography (CT) images (Barber et al. 2000). All machine learning operations are performed at the headquarters of the company.

The AI system was implemented almost two years prior to the study. It was first introduced based on a clinical trial and some radiologists evaluated the accuracy and functionalities of this system through studies. Then, a small group of radiologists received a dedicated training in the usage of the AI in which the possibilities of AI errors were discussed. Later, the tool was implemented into clinical routine for diagnosing stroke and used by almost all radiologists on a regular basis. The performance of the AI system was assessed in clinical studies and found to be similar to the performance of stroke experts with a specificity within a range of 90 to 95% and a sensitivity of around 50% (Herweh et al. 2016).

In simple terms, the general process of diagnosing stroke in the hospital consisted of two steps. First, radiologists made a binary assessment whether the patient acutely suffered a stroke or not. This process needs to happen fast, because timely treatment is essential to save a stroke patient's life. Second, radiologists formed a differentiated diagnosis and created a detailed report about the patient. During this process, they had more time and could consider more details. Experienced radiologists often develop the first estimate within seconds. Novice radiologists need substantial training before they obtain the capability to read and interpret images.

In the case hospital, most of the novices and expert physicians conducted a first initial assessment by using the original computed tomography images (native CT image) and without the support of the AI system. Even though the analysis of the AI system was provided fast, multiple interviewed radiologists reported that it still took too long and that they needed to decide faster in those situations. Thus, the investigated AI system was mostly used in the second phase of the decision making process, in which a detailed report about the patient was created and a differentiated diagnosis was made to quantify the severity of the stroke and assess which brain areas were damaged.

During this process, the radiologists in the University Hospital could assess different pieces of information using the AI system: First, they could utilize the ASPECT score, which is a quantification of the brain areas that are damaged. The score ranges from 0 to 10, with 10 meaning acute stroke with all brain areas affected, whereas lower numbers quantify different damaged brain areas. We refer to this as *score* during the data analysis. From a conceptual perspective, the score can be considered as *binary advice* if radiologists only utilize it to assess whether the patient suffers from a stroke or not. Second, they could assess a detailed, colored computed tomography image, in which the AI system showed which segments of the brain were classified as critical and which were classified as unharmed. This image served as a visualization and quantified the damages in more detailed areas; we refer to it as *quantification*. As the initial diagnosis had already been developed with the help of the native CT in the first step, radiologists could then compare their or their colleagues' initial judgment with the AI systems output. Yet, the radiologists differed considerably in how they considered the provided information of the AI system, and in which sequence they assessed the native CT image, the score, and the quantification.

3.2 Data Collection

One of the authors and a research assistant spent two days at the radiology department of the University Hospital in fall 2017. We observed how radiologists interacted with the system throughout their workdays and how they included AI advice in their diagnostics decisions. We collected 14 in-depth semi-structured interviews with chief, senior, and assistant physicians, each one lasting between 30 min to one hour. We sampled participants through personal referral from one radiologist to the next. Further, we collected background material and documents, such as publications concerning the AI system, and conducted one interview with the company developing the AI system to better understand its functionalities. The two researchers kept notes of each observation and later, in the evenings, discussed the findings with a researcher who did not participate in the data collection.

Our data collection already accounted for the two relevant insights from prior work that emphasized on the importance of confirmation versus disconfirmation (Jussupow et al. 2021) and of how confident a decision maker is about their ability to make a correct decision without a supporting information system (Burton et al. 2020). Nonetheless, we always remained flexible to incorporate new elements and ideas that came up during data collection such as usage patterns that describe different ways of using the diagnostic AI system in clinical practice. All participants were interviewed during clinical routine and described their experience with the AI system based on a recently assessed patient case and by demonstrating the usage of the AI system on the computer. All participants voluntarily talked about situations in which the system confirmed

them. With the help of an interview guide (Appendix A, available online via http://link.springer.de), we additionally engaged participants in discussions about situations in which the system disconfirmed them. When participants stated their reactions to these situations, we followed-up with ad-hoc personalized questions to uncover their reasoning, motivations, and potential consequences for their usage patterns. Aiming to sample participants with diverse diagnostic self-efficacy, we acquired participants of different levels of work experience and expertise. As a first means of shedding some light on each participant's perceived diagnostic self-efficacy, we also used three survey items as well as several questions about their experience, how their experience related to stroke diagnostics and AI, as well as their overall work situation (Appendix A). To avoid the effect of possible desirability bias, we compared the three-item survey with other qualitative responses of each interviewed radiologist.

Drawing on all these inputs, the two researchers present at the case site then classified each interviewee independently as having low, medium, or high diagnostic self-efficacy. A third researcher double-checked these classifications and helped to resolve disagreement about the assessment of one interviewee. Importantly, the qualitative assessment by the researchers could differ from participants' self-report in the survey items. For example, participant #13 answered the quantitative survey questions with 7/10 points of confidence, but qualitatively amended the rating with statements about seeing this assessment in light of other assistant physicians' (lack of) skills, being "a young assistant physician only," and about being "rather pessimistic" about diagnosis quality in general. Consequently, participant #13 was assessed as having low diagnostic self-efficacy despite a medium self-reported survey rank. Table 1 provides an overview of the study participants and shows that all levels of diagnostic self-efficacy are present in our sample.

3.3 Data Analysis

For data analysis, we borrowed analytical devices from the grounded theory methodology but adapted them to our needs in order to account for pre-existing knowledge and concepts. Specifically, we relied on pre-existing concepts of confirmation and disconfirmation (Table 2) to structure our analysis of radiologists' interactions with the diagnostic AI system, making sure to attend to differences between confirming and disconfirming AI advice. We also assessed diagnostic confidence as a potentially meaningful factor for different usage patterns of the diagnostic AI system. However, we realized during data analysis[1] that diagnostic self-efficacy was the more meaningful concept for understanding usage patterns across multiple diagnostic decisions whereas diagnostic confidence helped understanding single decisions. We engaged in descriptive, axial and selective coding (Charmaz 2006; Saldaña 2013) using NVivo for written codes and memos as well as whiteboards for drawings and visualizations. Descriptive codes consisted of the pre-existing concepts as well as open codes referring to different narratives of evaluations in decision making tasks, perspectives on radiologists' relation to increasingly powerful AI systems, and expectations towards AI systems in the future of clinical practice. In line with recommendations (Saldaña 2013), descriptive codes could still have strong conceptual overlap.

The codes were then refined and iteratively aggregated. For example, several evaluations of confirming AI advice (Table 3) were iteratively aggregated to the concept *Bolstering diagnostic confidence*. Axial codes were then used to describe emergent relationships between core concepts. Specifically, we realized that there were different associations between evaluations of confirming and disconfirming AI, elaborations on how frequent the AI system was utilized and different levels of diagnostic self-efficacy. This resulted in the development and characterization of three usage patterns. Each participant was classified into one dominant usage pattern. We used written memos and diagrams to discuss among the author team how we reasoned that the different participants developed their usage pattern, constantly comparing our reasoning to the data and patterns in the data to each other. Once the researchers involved in data analysis had reached agreement on the central developments, the third author of this paper critically double checked the reasoning, requesting clarification and evidence from the data where necessary.

Finally, we engaged in selective coding that elaborated the preliminary theoretical mechanisms underlying the patterns in the data. In an iterative process of reasoning from the data, comparing results to the literature, and refining our interpretations, we arrived at three organizational sensemaking processes. Those processes helped us to theorize why different evaluations of confirming and disconfirming AI advice are closely interrelated, whereas others almost never occur in combination and how this results in different usage patterns. Furthermore, the theoretical lens allowed us to interpret the narratives of the interviewed radiologists in more detail: For example, a senior physician started the interview by clarifying which types of mistakes the AI system typically performs and how those mistakes disconfirm the assessment of a radiologist, making it necessary that radiologists actively engage in the evaluation of the system. After talking to this radiologist in detail about his own usage pattern, he

[1] We thank the associate editor and anonymous reviewers for pointing us in this valuable direction.

Table 1 Overview of study participants

Number	Medical role within radiology department	Years of experience	Mean of self-reported diagnostic confidence [1–10]	Classified diagnostic self-efficacy	AI skills
1	Chief	23	Missing	High	Has experience with AI systems in different contexts
2	Senior	15	10.0	High	Researches AI systems
3	Chief	12	9.0	High	Leader of an AI group, research on AI systems
4	Senior	10.5	6.3	Medium	–
5	Senior	7.5	8.0	Medium	–
6	Assistant	7	9.3	High	–
7	Senior	7	7.0	Medium	Develops AI systems
8	Assistant	5.5	8.0	Medium	–
9	Assistant	5.5	6.2	Low	–
10	Assistant	5	8.0	Medium	–
11	Assistant	2.5	7.7	Medium	–
12	Assistant	2.5	5.0	Low	–
13	Assistant	2.5	7.0	Low	–
14	Assistant	0.25	4.3	Low	Develops AI systems

Table 2 Overview of concepts

Concept	Definition (Jussupow et al. 2021)
Pre-existing concepts	
Confirmation in decision task	AI system matches one's own judgment that was formed through evaluation of the clinical information (Jussupow et al. 2021)
Disconfirmation in decision task	AI system conflicts with one's own judgment or other information (Jussupow et al. 2021)
Diagnostic confidence	A diagnostic decision maker's perception on how certain they are to make the right decision drawing on their own analysis of a specific decision task (Jussupow et al. 2021)
Emergent and refined concepts (see also Tables 3 and 4)	
Diagnostic self-efficacy	A medical decision maker's belief in their capabilities to make correct diagnostic decisions (adapted from Bandura 1997)
Intensifying usage	Engaging intensively with the AI system and using it extensively during clinical routine
Deflecting usage	Engaging superficially with the AI system, but suggesting that others use it extensively
Abandoning usage	Engaging minimally with the system and suggesting to remove it from clinical practice
Sensedemanding	Acquire and process information to ameliorate uncertainty and equivocality (Gioia and Chittipeddi 1991)
Sensegiving	Influence the sensemaking of others toward a preferred interpretation (Vlaar et al. 2008)
Sensebreaking	Break down or destroy meaning to induce new ways of thinking and acting (Tan et al. 2020)

described that he tended to probe the AI and to test its accuracy, but that disconfirming AI advice did not affect his judgment. With the help of the sensemaking lens, we were thereby able to classify this narrative as a *sensegiving* process with the goal to disseminate knowledge about the AI accuracy, which resulted in the deflecting usage pattern displayed by this radiologist in the own decision making.

Tables 2, 3 and 4 depict the core concepts we retained after this process.

4 Results

The goal of this paper is to understand how radiologists differ in their usage patterns of diagnostic AI systems and

Table 3 Overview of emerging confirmation codes in clinical practice

First-order codes of evaluating confirming AI advice			Emerging second-order categories
Descriptive code	Exemplary quote	Information usage	Description
1. Using as a second opinion	"Because you can have (the AI system) as a control mechanism." (Assistant physician #12)	Binary evaluation (acute stroke?)	Bolstering own diagnostic confidence: Confirmation by AI system increases radiologist's diagnostic confidence in a diagnostic decision
	"I use this system as a support to check the patient case a second time retrospectively (…) OK, the computer program has the same opinion as I have" (Assistant physician #9)	Binary evaluation & detailed assessment	
2. Deciding between conflicting options	"(…) if I am unsure regarding two possible options and SYSTEM confirms me in one of them, then I would rather go with that one" (Assistant physician #11)	Detailed assessment	
3. Justifying clinical communication	"(…) but for the classification of the extent SYSTEM was still a help and (I could then) tell or show the clinician (that) I am not the only one who sees this here, but the system has recognized it as well." (Assistant physician #10)	Binary evaluation & detailed assessment	
4. Checking plausibility	"I usually take a brief look at it and think about whether it is plausible, whether it fits to the clinical information. (…) And there will be more situations in which the radiologist is a plausibility checker (for the AI advice)." (Chief physician #3)	Binary evaluation & detailed assessment	Probing the system: Radiologist checks whether system is able to come to the same result; Radiologist judges confirmation as an indication of system accuracy
5. Competing against the system	"Currently, I am looking at the system with my left eye only, just to check whether it is able to match my own judgment." (Senior physician #2)	Detailed assessment	

what role confirmation and disconfirmation play in forming these patterns. We first introduce the identified usage patterns. Then, we describe how these usage patterns relate to confirmation and disconfirmation, whilst theorizing how usage patterns develop through distinct sensemaking activities based on radiologists' diagnostic self-efficacy.

4.1 Emerging Usage Patterns in Clinical Practice

Three distinct usage patterns emerged from the data: Intensifying usage, deflecting usage, and abandoning the system. These usage patterns constituted the results of a longer process and have been observed at the point of data collection, two years after the AI system had been introduced and established in clinical practice at the case site. All radiologists in our study showed one of the three usage patterns.

4.1.1 Intensifying Usage

Radiologists exhibiting the first usage pattern engaged with the AI system with increasing intensity and applied it extensively during clinical routine. These radiologists saw benefits in the information provided by the system for both making a first assessment about whether the patient was acutely suffering from a stroke and for preparing a detailed report with a diagnosis of the different brain areas affected. Being convinced of the system's usefulness, they used the AI system routinely in all their stroke diagnosis tasks. All radiologists exhibiting this pattern emphasized on the system's role as a backup that made sure they did not overlook any critical fact, even under high pressure and in stressful situations. For example, assistant physician #13 stated:

> "(It) is just when you are wrung out at four in the morning again and you just slept an hour; the eye is not yet so awake; then you simply have a second opinion (of the AI system) that confirms (your own assessment)." (Assistant physician #13).

Importantly, none of the radiologists were relying on the AI system as the primary means of diagnosing the patient case. Instead, they all had the skills and knowledge to assess patient cases independently and did so following clinical routine. Especially in situations where fast treatment for acute stroke was potentially necessary, the system was used to make quick clinical decisions when it

Table 4 Overview of emerging disconfirmation codes in clinical practice

First-order codes of evaluating disconfirming AI advice			Emerging second-order categories
Descriptive code	Exemplary quote	Information usage	Description
1. Remaining open for feedback	"So, let's say I don't see anything on the native CT image, but the system recognizes something on one side and the side also matches with the clinic…. Yes, then I would reconsider my decision and perhaps even do further diagnostics" (Assistant physician #10)	Binary evaluation; Acquire additional clinical information	Compensating for mistakes: Mechanisms that help radiologists to evaluate the disconfirming AI advice and identify the error, either in their own or in the system's assessment
2. Following clinical routines	"(…) so, if it (the system) deviates (from my own view), then one checks first if this different judgment would result in a different treatment for the patient. And if I think that I am unsure I can always call a senior physician." (Assistant physician #13)	Detailed assessment; Use information according to protocol	
3. Knowing about common AI errors	"For example, the system highlights the wrong side (of the brain) or classifies something (an infarct) that is clearly old as acute. Those have always been obvious errors." (Senior physician #5)	Detailed assessment; Targeted comparison with clinical information	
4. Ignoring disconfirming AI advice	(Interviewer): "And if this system says something else than you would have expected. How do you react then?" – Senior physician #5: "Then we ignore it"	Binary evaluation; Brief comparison	Rejecting without detailed consideration: Fast default response to disconfirming advice
5. Feeling irritated	"The system causes irritation. (..) One only gets (unnecessarily) confused." (Senior physician #4)	Score and detailed assessment; refuse to examine additional information	Focusing on AI errors: Reject AI advice after brief doubts; emphasizing on situations in which the system failed to make correct assessments before
6. Playing down system accuracy	Interviewee: "There is a study from our department, I think they found out how accurate it (the AI system) is." Interviewer: "I think it was about 87% accurate. That is rather high, isn't it?" Interviewee: "Anyway, I didn't find it reliable enough in clinical routine." (Assistant physician #6)		

confirmed the radiologist's first impressions. For example, assistant physician #12 reported:

> "So, when I don't see anything critical and I see 10 out of 10 points on the (AI system) score, I don't go into all the details of the images. I quickly go over them to make sure there is no mismatch."

Later, for creating a detailed diagnostic report, these physicians used the information provided by the AI system in an iterative process to dig deeper into potential damages in different brain areas. The detailed, quantified views on different segments of the patient's brain were seen as especially helpful to decide between alternatives and in case of boundary decisions when classifying the severity of the stroke. Nonetheless, as all other radiologists, those who exhibited intensifying usage were aware that the AI system was not perfect. While a clinical study at the case hospital had shown that the AI system had an overall accuracy rate of about 90%, the radiologists mentioned that the system did make some erroneous assessments. However, radiologists who intensified their usage perceived that the benefits clearly outweighed any drawbacks. Assistant physician #10 put it as follows:

> "Sometimes the system classifies it (an infarct) as an old one or does not recognize it as acute. So, there are still some errors. (...) But often it (the severity of a stroke) is about a tiny (brain) area. And if the system recognizes this area reliably, that is a huge help."

4.1.2 Deflecting Usage

Radiologists following the deflecting usage pattern considered the AI system as helpful and supportive, however not for themselves but for less experienced colleagues. These radiologists used the AI system superficially during clinical practice, but the system rarely had any effect on their actual diagnosis decisions. In fact, several radiologists in this group stated that the system did not influence their

decisions at all. Nonetheless, and despite a high but imperfect system accuracy, these radiologists were convinced of the system's usefulness for less expert colleagues. Chief physician #3 said:

> "I mean, I am chief physician in this area. I have seen many strokes in native CT and CT-angiography. I don't rely on it (the system), and I don't allow it to influence my decision. (...) A youngster in his first year would say 'Oh, I have to look at this in more detail!' And I think that is what the system can do; it gives you a nudge in the right direction. And if you are learning and take the clinical information into account this can lead to a strong diagnostic procedure."

4.1.3 Abandoning the System

Finally, radiologists following the abandoning pattern engaged minimally with the AI system and suggested removing it from clinical practice. Typically, they stopped actively using it or conducted a bare minimum of interactions with the system. A senior physician from this group reported being highly convinced of the AI system's accuracy at its introduction but gradually, with more clinical usage, becoming disappointed with its accuracy and perceiving it as not sensitive enough. The radiologists abandoning the system did not see mentionable value of the AI system, neither for themselves nor for their colleagues. Assistant physician #6 put it as follows:

> "I've been practicing radiology for the last seven years, and most of that time I've been practicing neuroradiology. I have a lot of experience in stroke imaging, especially with CT and MRT. I dare say I can make a (proper) diagnosis. And therefore, I do not need feedback from a software of which I think it is rarely right."

Radiologists abandoning the system did not record the system's assessments in their detailed diagnostic reports and did not rely on its binary assessments for making decisions early on. Instead, these radiologists highlighted the importance of personal skills, experience, and competence as decisive for proper diagnostics. Overall, radiologists exhibiting this pattern did not see benefits of the system.

4.2 AI Advice in Decision Making

The interviewed radiologists elaborated intensively on the general influence of the diagnostic AI system on their decision making, often through back-and-forth reflections of confirming and disconfirming interactions with the AI system. Out of 14 interviews, we found eight specific diagnostic examples with a specific patient case, while the other radiologists described more their general approach in the interaction with the AI system. Interestingly, those specific examples included mostly disconfirmation through mistakes of the AI system, while the general description consisted mostly of confirmation events. In the following, we outline different patterns how confirming and disconfirming AI advice were evaluated by the interviewed radiologists.

4.2.1 Evaluation of Confirming AI Advice

The interviews indicated that radiologists at the case site not only differed in their usage patterns in terms of how intensively they used the AI system; they also described different effects of AI advice on their diagnostic decision making. As a result, the radiologists described different narratives when the AI system confirmed or disconfirmed their own assessment of a patient case. Tables 3 and 4 show excerpts from our data analysis, displaying exemplary, openly coded descriptive codes and their aggregation to emerging, mutually exclusive second-order categories. These emergent second-order categories laid the foundation for better understanding radiologists' sensemaking processes of diagnostic AI advice.

Multiple radiologists described clinical decisions in which the AI system had confirmed their own diagnostic reasoning and *bolstered their diagnostic confidence* (Table 3). During both, the urgent, binary evaluation whether a patient was acutely suffering from a stroke and the subsequent detailed diagnostics, some radiologists appreciated the information provided by the system if it confirmed their own assessment of the patient case. The ASPECT score was seen as a quickly available piece of advice about whether a patient suffered an acute stroke, allowing faster decision making in this regard. The quantification of image data and visualizations were perceived as useful for making the detailed assessment.

For example, radiologists referred to the system as a control mechanism and as an immediately available second opinion (Table 3). They perceived this as beneficial because it allowed them to be more confident in making their diagnostic decisions, knowing that their own assessments and the AI advice were well aligned, even when they were working under stressful conditions such as sleep deprivation. In situations that required assessing the damage to different brain areas in detail for further treatment, several radiologists argued that the system helped them to more confidently make boundary decisions, for example, to classify damage severity when other indicators allowed ambiguous interpretations. In the same way, some radiologists argued that they could more easily justify and

communicate their assessment to colleagues, and sometimes even to the patient, if they could rely on the systems' confirmation of their own evaluation (Table 3). Overall, confirming advice of the AI system bolstered these radiologists' diagnostic confidence and was seen as very desirable.

Other radiologists used the system primarily to *probe it* rather than aiming to gain input for better decision making (Table 3). These radiologists stressed the need to constantly evaluate the accuracy of the system advice by checking its plausibility considering available clinical data, suggesting that human radiologists are generally better at accounting for additional clinical data than the AI system. Some radiologists stylized their evaluation of the AI advice to a competition, insinuating a sense of rivalry between human and machine about making the better diagnostic assessment. When the AI system came to the same conclusions as these radiologists, they typically saw this as a positive indicator of system accuracy (Table 3). In sum, radiologists who saw benefits in using the AI system, evaluated confirming AI advice either as a means of bolstering their diagnostic confidence and making better diagnostic decisions or as a way of probing the system's accuracy.

4.2.2 Evaluation of Disconfirming AI Advice

The radiologists also faced decisions in which the AI system disconfirmed their own diagnostic reasoning in clinical practice and described how they reacted. Overall, each radiologist adhered to one of three strategies when dealing with disconfirming AI advice (Table 4). Some of the radiologists engaged with the system to *compensate for mistakes*. These radiologists had accumulated knowledge and practices that aimed to ensure that no erroneous assessments would impact the treatment of their patients, neither errors produced by the system nor errors produced by themselves. For example, radiologists intended to stay open for feedback, even during urgent binary evaluation of whether a patient acutely suffered from a stroke. Especially if some pieces of clinical information could be interpreted in the same way as the system's assessment, these radiologists would reconsider their initial judgment and critically reflect on what additional information they needed to make a definite decision (Table 4). In doing so, many radiologists perceived adhering to clinical routines and protocols as crucial, some of which had been adapted to the use of the system. Assistant physician #8 put it as follows:

> "We have a clear algorithm how to proceed. You go for a native skull CT and then, depending on the clinical setting, also perfusion CT and angio CT. (...) And if you still feel uncertain you will go further steps. (...) And in that way, (the system's) ASPECT score is a usable and reasonable puzzle piece in the initial diagnostics."

Moreover, those radiologists who used the system for compensating for mistakes also aimed to create knowledge about common AI errors (Table 4). For example, it was well known to the radiologists that the system sometimes classified old infarcts in brain tissue as acute ones. Although such a conclusion could be reached by examining only the CT images, clinical information about patient behaviors and perceptions that were collected during standard procedures for all potential stroke patients could quickly rule it out. Similarly, specific ranges of grey values in CT images had been identified as a potential cause of erroneous AI assessments. Knowing about this limitation, the radiologists double-checked all system assessments that were based on image areas containing these grey values. Overall, practices and knowledge helped these radiologists to compensate for potential errors of either side.

A second strategy observed in several radiologists during the time-critical initial binary evaluation was that of *rejecting without detailed consideration*, describing a rejection of disconfirming AI advice without further considering it because it was seen as unlikely that the system was correct in light of the radiologist's prior assessment (Table 4). Radiologists following this strategy assumed that the relatively few cases in which the system disconfirmed their own judgment must be due to errors of the software. Those radiologists had detailed knowledge about the contexts of typical AI errors and used this knowledge without further considering the details of the case. Considering the time pressure and potential negative effects on patient health that could be caused by delayed decisions, these radiologists decided actively to ignore disconfirming AI advice and rather make a quick decision based on their own assessment.

Lastly, a third strategy we labeled *focus on errors* described some radiologists' complete rejection of disconfirming AI advice both during the initial binary evaluation as well as during subsequent detailed assessments while they referred to system errors they had previously observed or thought to have observed (Table 4). Disconfirming advice by the AI system was perceived as irritating and as an unnecessary source of distraction. In contrast to their colleagues who only engaged in fast rejection during the time-critical binary evaluation, these radiologists did not reject the system advice because of time pressure but rather argued that it was near to impossible that the system could outperform a human expert stroke assessor. Even confronted with objective measurements of high system accuracy from their own department, they emphasized their negative experiences with the system and suggested not to rely on it.

Constant comparison of commonalities and differences between the radiologists who exhibited each usage pattern showed an association between emerging usage patterns, how radiologists described their diagnostic evaluation of confirming and disconfirming AI advice, and their diagnostic self-efficacy (see Table 5). Radiologists who were classified as displaying intensive usage engaged with confirming AI advice to bolster their diagnostic confidence and with disconfirming AI advice to compensate for mistakes. These radiologists had low to medium diagnostic self-efficacy. Radiologists who showed deflecting usage behavior engaged with the confirming system in probing its accuracy and by quickly rejecting disconfirming AI assessments. This association was found in radiologists with medium to high diagnostic self-efficacy. Finally, radiologists showing abandoning usage behavior paid minimal or no attention to confirming AI advice and focused on errors when confronted with disconfirming AI assessments. This association occurred more frequently in radiologists with high diagnostic self-efficacy.

4.3 Analytical Abstraction: Usage Patterns as a Result of Sensemaking Processes

After carefully analyzing our results, we realized that the three observed usage patterns constituted the results of three sensemaking processes revolving around the radiologists' interpretation of confirmation and disconfirmation by the AI system in light of their diagnostic self-efficacy. Figure 1 depicts the three identified sensemaking processes.

The first process describes how *sensedemanding* can explain the intensifying usage pattern. In our sample, many radiologists with low and medium diagnostic self-efficacy engaged in sensedemanding. They intensively engaged with the system assessments—with confirming ones as well as with disconfirming ones—in a way that reduced uncertainty and equivocality in diagnostic situations. Rooted in low to medium diagnostic self-efficacy, these radiologists did not enter diagnostic decision tasks with many preconceptions; they rather absorbed the information provided by the AI in each case to build diagnostic confidence and develop a diagnostic decision. Those radiologists evaluated confirming AI advice in a way that bolstered their diagnostic confidence, increasing their diagnostic self-efficacy across multiple diagnostic decisions. Therefore, radiologists who engaged in sensedemanding proactively sought confirmation by the AI system in situations in which they were not confident enough. If the AI system disconfirmed their initial diagnostic assessment, radiologists who engaged in sensedemanding carefully evaluated their own reasoning and the AI system to compensate for mistakes of either. They asked for additional information about typical mistakes of the AI and carefully crosschecked the provided information by the AI with their own assessment. Sensedemanding enabled these radiologists to evaluate disconfirming AI assessments as a learning opportunity, either about what they had missed in assessing the patient case before receiving the system support or about situational weaknesses of the system and the conditions under which the system may not perform well. For example, assistant physician #10 said: "I have to understand what the system says… I will not just write down the number if I do not understand why" (Assistant physician #10). Further, we observed that radiologists described both evaluating confirming and disconfirming AI assessments as mutually reinforcing: On the one hand, the perceived benefits of confirming AI advice increased radiologists' willingness to engage more intensively with disconfirming AI advice, even if it means that the AI system had made a mistake. On the other hand, these radiologists experienced disconfirming AI advice as helpful for identifying mistakes in their own diagnostic reasoning, which further increased the engagement with the AI system in diagnostic situations. Motivated by sensedemading, radiologists intensify their usage of the AI system. Based on those observed processes we propose the following relationships:

> **P1:** Radiologists with lower levels of diagnostic self-efficacy are more likely to engage in sensedemanding with diagnostic AI systems than radiologists with higher levels of diagnostic self-efficacy.

Table 5 Associations between AI usage patterns and radiologists' diagnostic self-efficacy

Emerging usage pattern	During confirmation	During disconfirmation	Diagnostic self-efficacy
1. Intensifying	Bolstering own diagnostic self-efficacy	Compensating for mistakes	Low Medium
2. Deflecting	Probing the system	Rejecting without detailed consideration	Medium High
3. Abandoning	Minimal or no engagement	Focusing on AI errors	High

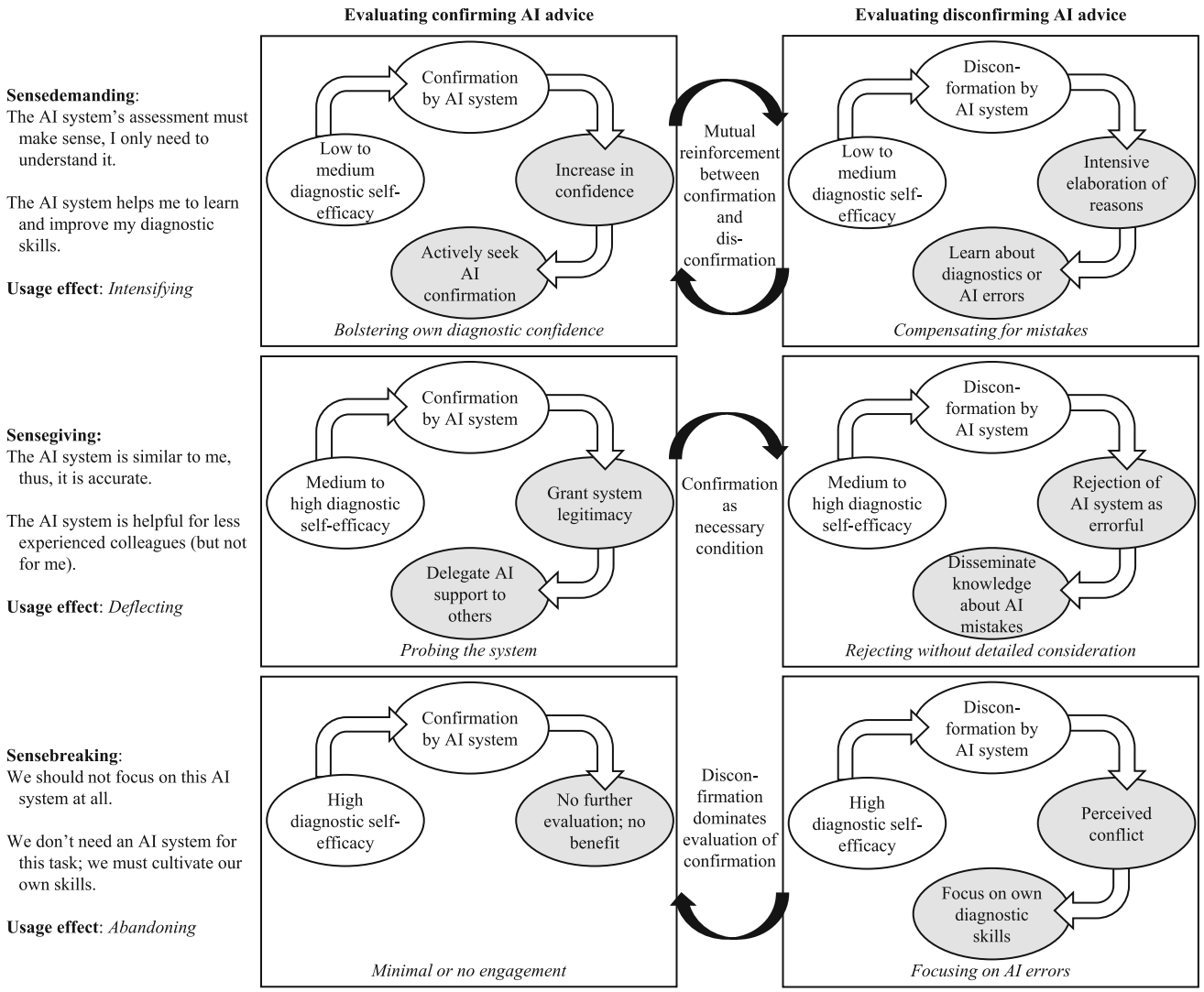

Fig. 1 Sensemaking processes during confirmation and disconfirmation from AI systems and effects on usage

P2: Over time, engaging in sensedemanding intensifies AI usage and increases radiologists' diagnostic self-efficacy.

The second process of making sense from AI systems presented in Fig. 1 is through *sensegiving*, whereby radiologists attempted to influence other radiologists' sensemaking activities toward a preferred interpretation of organizational reality. In our sample, radiologists' deflecting usage entailed the message that the AI system was indeed useful and should be used not by themselves but by less skilled and less experienced colleagues. With confirming AI assessment, those radiologists probed the system, sometimes even insinuating mock rivalry and competition between themselves and the AI system. With relatively high diagnostic self-efficacy, those radiologists demonstrated in clinical practice that they saw a basis for comparing the system's diagnostic capabilities to their own. These radiologists used patient cases in which the system came to the same conclusions as they did to openly grant the system legitimacy by showing that it was able to match them to some degree. On the other hand, these radiologists did not accept any real interference with their own diagnostic decision making by disconfirming AI assessments. They reported that disconfirming AI assessments had no effect on their own diagnostic decision making as those were swiftly ignored without detailed consideration. Instead, disconfirming AI assessments were used in a sensegiving way to reflect on typical mistakes of such systems in clinical situations. Radiologists who engaged in sensegiving aimed to ensure that those typical mistakes of AI systems were disseminated across the University Hospital and carefully integrated into the diagnostic reasoning of less experienced colleagues. In so doing, these radiologists actively played the part of gatekeepers who critically evaluated system performance and

served as a basis for comparison when assessing the AI system's adequacy for clinical practice. Nevertheless, evaluating confirming AI advice serves as a necessary condition for the sensegiving process as establishing the system's legitimacy is required before radiologists can effectively deflect usage. At the same time, frequent confirmation reinforced the individuals' feeling that the system did not provide diagnostic insight beyond what they themselves could achieve. Moreover, although these physicians actively used the system in clinical practice, they did not allow that the system influenced their decisions. Hence, using the AI system would not affect the diagnostic self-efficacy of those radiologists. Consequently, AI advice would not further influence diagnostic self-efficacy. Based on our findings, we make the following propositions:

P3: Radiologists with higher levels of diagnostic self-efficacy are more likely to engage in sensegiving with diagnostic AI systems than radiologists with lower levels of diagnostic self-efficacy.
P4: Over time radiologists who engage in sensegiving do not benefit in their diagnostic self-efficacy from AI systems and deflect the AI system usage to less experienced colleagues.

Finally, the third way of sensemaking as depicted in Fig. 1 is *sensebreaking* defined as attempts to break and destroy meaning in order to induce new ways of thinking. Specifically, radiologists who engaged in abandoning usage did not subdue to the thinking that they should be supported by AI systems that provided them with advice. Instead, these radiologists aimed to completely abandon the system and refocus the discussion on the human experts' diagnostic skills. As such, they saw no benefit in confirmatory AI advice as they had a high diagnostic self-efficacy and were already relatively confident in the correctness of their own judgment in specific decision situations. In situations in which the system disconfirmed them, they engaged in the focusing on error evaluation pattern. This resulted in a conflict between their expert opinion and the AI assessment, resulting in a feeling of irritation and annoyance, especially if the AI system made repeatedly the same mistake.

In contrast to radiologists who exhibited sensedemanding, radiologists applying sensebreaking did not aim to understand origins and boundary conditions of AI errors. Instead, they emphasized examples of obvious errors that the system had produced, denying the system any meaningful diagnostic capabilities, and described situations that highlighted the inferiority of the AI system in comparison to the human judgement. Hence, radiologists' narratives were centered around evaluations of disconfirming AI advice, dominating evaluations of confirming AI advice. In focusing on system errors, these radiologists intended to break the discussion about the system and to redirect it toward professional human intuition and agency. For example, the experienced assistant physician #6 stated:

"Also, the decision about thrombolysis is made based on gut feeling, taking time (since the stroke) and the visual how much of the (brain) territory has been damaged into account. Exactly. So, this 'SYSTEM', I don't use it. In the clinical routine, it has no relevance whatsoever." (Assistant physician #6).

Thus, abandoning the system not only constituted a mechanism for those radiologists to prevent the system from potentially challenging their high diagnostic self-efficacy, but it also allowed them to drive the discourse about better diagnostic decisions toward areas of human skills and tacit knowledge that were hard to ascribe to a technological solution. However, using the diagnostic AI system would not affect their diagnostic self-efficacy. Based on our findings, we propose:

P5: Radiologists with high levels of diagnostic self-efficacy are more likely to engage in sensebreaking than radiologists with low and medium levels of diagnostic self-efficacy.
P6: Over time, radiologists who engage in sensebreaking do not benefit in their diagnostic self-efficacy from the support of AI systems. Instead, they abandon using the AI system.

5 Discussion

This study set out to provide first insights into how radiologists differ in their usage of AI systems. Drawing on rich data from a revelatory case study, we elaborated three distinct processes of sensemaking from confirming and disconfirming AI assessments through which radiologists come to differ in their AI system usage. We developed six propositions on the role of diagnostic self-efficacy as an antecedent for sensemaking and on how, over time, diagnostic self-efficacy could change through sensemaking. Our preliminary account of sensemaking and usage of diagnostic AI systems in clinical practice paves the way for future research.

5.1 Contributions

With our findings, we offer multiple contributions to research on the usage of AI systems and their impact on knowledge work and on medical decision making. First, we elaborate three distinct sensemaking processes and resulting usage patterns in physicians. Whereas prior work has shown that confirmation and disconfirmation can trigger

different cognitions that decide about whether radiologists accept AI advice in isolated decision tasks (Jussupow et al. 2021), this study elaborated how radiologists make sense of confirmation and disconfirmation to form emergent usage patterns. While sensedemanding, sensegiving, and sensebreaking have been observed in other contexts before (Tan et al. 2020; Vlaar et al. 2008), this study provides a detailed and contextualized view of the three sensemaking processes in medical decision making with the support of AI systems. This generates a first understanding of how radiologists form their AI system usage patterns for complex decisions in which the factors that determine the system's advice are unknown and it is challenging to determine which decision is correct.

Second, this study extends research on decision making with AI systems that has emphasized the importance of decision makers' confidence (Dietvorst et al. 2015; Jussupow et al. 2021). This study elaborates the role of diagnostic self-efficacy as a central factor in three distinct sensemaking processes and resulting usage patterns: In sensegiving processes, diagnostic self-efficacy serves primarily as a potential source of legitimacy that can be granted to the system, but is not affected by the usage of the diagnostic AI system; in sensebreaking processes, radiologists focus on their own diagnostic skills and abandon using the diagnostic AI; in sensedemanding processes, diagnostic AI advice could actually alter radiologists' diagnostic self-efficacy over time. These insights elevate prior knowledge on confidence in isolated decision tasks to the level of emergent usage. Our qualitative study thereby helps to understand the implicit and contextually embedded processes that determine why radiologists show specific usage patterns that may not be understood by looking at isolated decision tasks in laboratory settings. It will, thus, be crucial to account for different sensemaking processes in clinical practice when moving the empirical research on decision tasks from laboratory experiments to the field. The results also call for future research on potential dynamic changes in the three sensemaking processes and their impact on diagnostic self-efficacy. As such, it is not unlikely that some radiologists can move from one sensemaking pattern to another over a longer period of time. For example, radiologists could primarily engage in sensedemanding when they first come to know about a diagnostic AI system but eventually engage more in sensebreaking after experiencing errors of the AI system. Moreover, the sensemaking process might dynamically change with the interaction context. For example, junior radiologists may change their sensemaking pattern of a diagnostic AI system when they are promoted to a new role that comes with altered job demands. Future work should, therefore, collect and investigate data on sensemaking from AI systems over longer periods of time and focus on how sensemaking processes change.

Third, this study provides insights for research on the usage consequences of AI errors and algorithm aversion (Baird and Maruping 2021; Berger et al. 2021; Burton et al. 2020; Dietvorst and Bharti 2020; Dietvorst et al. 2015; Jussupow et al. 2020; Longoni et al. 2019). The findings suggest that individuals' professional contexts have a major influence on how they deal with errors of AI systems. In our study, there was only one system applied to very similar tasks by all decision makers. Thus, the errors that were observed were arguably also very similar in type and frequency. However, reactions differed vastly. Individuals with lower diagnostic self-efficacy often engaged in sensedemanding, tried to understand the origins of different errors of the AI system, and embedded this knowledge about AI errors into their diagnostic practices. Individuals with high diagnostic self-efficacy more often engaged in sensebreaking and denied any meaningful diagnostic system capabilities by pointing to system errors, even though studies in their own department suggested rather high accuracy. Thus, occasional errors were actually acceptable for sensedemanding individuals, if they could understand when to expect them, and did not hamper intensified usage at all, whereas the detail, origin, and frequency of errors did not actually influence the rejection and abandoning of the system by sensebreaking individuals. Thus, future research on algorithm aversion may have to account very thoroughly for individual context, for knowledge about AI errors, and for sensemaking processes that span more than only few decision tasks.

5.2 Limitations

Our study bears some limitations that need to be considered for interpreting the results. First, from a methodological perspective, we considered only one case site in Germany and investigated the usage of one specific diagnostic AI system. Hence, it is unclear how often different patterns occur in other sites and whether more or other patterns would emerge in different settings. Moreover, our case setting does not actually allow us to tell which sensemaking processes are more effective or lead to better decisions than others. In particular, we do not know whether the actual decisions that radiologists make, for example, if they ignore the AI system, result in more accurate decision outcomes. In the tradition of revelatory case studies (Sarker et al. 2012), we believe nonetheless that our findings can help understand similar settings; namely, situations in which physicians must regularly make complex diagnostic decisions using black-box AI systems that have a comparable level of accuracy.

Second, from a conceptual perspective, there are several limitations in our classification of diagnostic interactions as confirmation and disconfirmation that can stimulate future research. On the one hand, we did not consider in detail whether the interactions described by the interviewed participants relate to false-negatives (underdiagnosing) or false-positives (overdiagnosing). Considering that the implemented diagnostic AI system displayed high specificity and a medium sensitivity, the most frequent errors were false-positives. Due to the limitations of interview reports and uncertainty regarding the correct final diagnosis in the clinical setting, it is difficult to observe the impact of those two different types of errors on diagnostic decisions. On the other hand, it is important to acknowledge the complexity of radiological diagnostic decision making. Trained radiologists are capable of concurrently assessing multiple diagnoses and differentiating between them by drawing on radiological images and the accompanying medical history of each patient case. To simplify our argumentation in our study, we conceptually synthesized this complex diagnostic reasoning into one diagnostic assessment that can be confirmed of disconfirmed by a binary advice of an AI system. However, if a radiologist has multiple hypotheses about possible diagnoses of a patient, it is difficult to conceptually distinguish confirmation from disconfirmation as one diagnostic hypothesis might be confirmed while the others are disconfirmed at the same time. Furthermore, due to hindsight and desirability biases, radiologists might be more likely to report their interactions with the AI system as confirming their own assessment than as disconfirming their own assessment which would have made them adjust the assessment. Future research should therefore conduct field experiments to investigate decision accuracy in combination with AI systems to assess the diagnostic decision making not in hindsight, but during its occurrence. Furthermore, more research is needed to understand how different pieces of information that are provided by the AI influence diagnostic decision making.

6 Conclusion

Overall, our study provides first insights into how radiologists utilize diagnostic AI systems in clinical practice. We elaborate that AI system usage can be considered as resulting from three distinct sensemaking patterns, which radiologists apply to assess diagnostic AI advice that confirms or disconfirms their initial assessments. Further, we show that diagnostic self-efficacy influences in which sensemaking process radiologists engage and how they evaluate errors of AI systems. We hope that our study serves as springboard for future research on the impact of AI systems on knowledge work and decision making.

Funding Open Access funding enabled and organized by Projekt DEAL.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12599-022-00750-2.

References

Arkes HR, Shaffer VA, Medow MA (2007) Patients derogate physicians who use a computer-assisted diagnostic aid. Med Decis Making 27(2):189–202

Baird A, Maruping LM (2021) The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. MIS Q 45(1):315–341

Bandura A (1997) Self-efficacy: the exercise of control. Freeman, New York

Barber PA, Bemchuk AM, Jinjin Z, Buchan AM (2000) Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. Lancet 355(9221):1670–1674

Berg M (1997) Rationalizing medical work: decision-support techniques and medical practices. In: Massachusetts Institute of Technology. MIT Press, New Baskerville

Berger B, Adam M, Rühr A, Benlian A (2021) Watch me improve—algorithm aversion and demonstrating the ability to learn. Bus Inf Syst Eng 63(1):55–68

Burton JW, Stein M, Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. J Behav Decis Making 23(2):220–239

Burton-Jones A, Volkoff O (2017) How can we develop contextualized theories of effective use? A demonstration in the context of community-care electronic health records. Inf Syst Res 28(3):468–489

Califf C, Sarker S, Sarker S (2020) The bright and dark sides of technostress: a mixed-methods study involving healthcare IT. MIS Q 44(2):809–856

Charmaz K (2006) Constructing grounded theory: a practical guide through qualitative research. Sage, London

Dietvorst BJ, Bharti S (2020) People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychol Sci 31(10):1302–1314

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126

Faraj S, Pachidi S, Sayegh K (2018) Working and organizing in the age of the learning algorithm. Inf Organ 28(1):62–70

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

Fügener A, Grahl J, Gupta A, Ketter W (2021) Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. Inf Syst Res. https://doi.org/10.1287/isre.2021.1079

Gioia DA, Chittipeddi K (1991) Sensemaking and sense-giving in strategic change initiation. Strateg Manag J 12(6):433–448

Glikson E, Woolley AW (2020) Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 14(2):627–660

Goodyear K, Parasuraman R, Chernyak S, Madhavan P, Deshpande G, Krueger F (2016) Advice taking from humans and machines: an fMRI and effective connectivity study. Front Hum Neurosci 10:542–557

Herweh C, Ringleb PA, Rauch G, Gerry S, Behrens L, Möhlenbruch M, Nagel S et al (2016) Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. Int J Stroke 11(4):438–445

Hinton G (2016) Geoff Hinton: on Radiology. https://www.youtube.com/watch?v=2HMPRXstSvQ&t=29s. Accessed 18 Mar 2021

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. Nat Rev Cancer 18(8):500–510

Hsieh J, Rai A, Xu SX (2011) Extracting business value from IT: a sensemaking perspective of post-adoptive use. Manage Sci 57(11):2018–2039

Jenkin TA, Chan YE, Sabherwal R (2019) Mutual understanding in information systems development: changes within and across projects. MIS Q 43(2):649–671

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

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision making process with artificial intelligence. Inf Syst Res 32(3):713–135

Jussupow E, Benbasat I, Heinzl A (2020) Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In: 28th European Conference on Information Systems, Marrakech. https://aisel.aisnet.org/ecis2020_rp/168

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

Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: the new contested terrain of control. Acad Manag Ann 14(1):366–410

Klein G, Phillips JK, Rall EL, Peluso D (2007) A data-frame theory of sensemaking. In: Expertise out of context: proceedings of the 6th international conference on naturalistic decision making, pp 113–155

Lapointe L, Rivard S (2005) A multilevel model of resistance to information technology implementation. MIS Q 29(3):461–491

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

Miller DD, Brown EW (2018) Artificial intelligence in medical practice: the question to the answer? Am J Med 131(2):129–133

Pachidi S, Berends H, Faraj S, Huysman M (2021) Make way for the algorithms: symbolic actions and change in a regime of knowing. Organ Sci 32(1):18–41

Petriglieri JL (2011) Under threat: responses to and the consequences of threats to individuals' identities. Acad Manag Rev 36(4):641–662

Pratt MG, Rockmann KW, Kaufmann JB (2006) Constructing professional identity: the role of work and identity learning cycles in the customization of identity among medical residents. Acad Manag J 49(2):235–262

Romanow D, Rai A, Keil M (2018) CPOE-enabled coordination: appropriation for deep structure use and impacts on patient outcomes. MIS Q 42(1):189–212

Russell SJ (2019) Human compatible: artificial intelligence and the problem of control. Viking, New York

Russell SJ, Norvig P (2010) Artificial intelligence: a modern approach, 3rd edn. Pearson, Essex

Saldaña J (2013) The coding manual for qualitative researchers. Sage, Thousand Oaks

Sarker S, Sarker S, Sahaym A, Bjørn-Andersen N (2012) Exploring value cocreation in relationships between an ERP vendor and its partners: a revelatory case study. MIS Q 36(1):317–338

Sarker S, Xiao X, Beaulieu T, Lee AS (2018) Learning from first-generation qualitative approaches in the IS discipline: an evolutionary view and some implications for authors and evaluators, part 2/2. J Assoc Inf Syst 19:752–774

Schuetz S, Venkatesh V (2020) Research perspectives: the rise of human machines: how cognitive computing systems challenge assumptions of user-system interaction. J Assoc Inf Syst 21(2):460–482

Shaffer VA, Probst CA, Merkle EC, Arkes HR, Medow MA (2013) Why do patients derogate physicians who use a computer-based diagnostic support system? Med Decis Making 33(1):108–118

Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, Ming W-K et al (2019) Artificial intelligence versus clinician in disease diagnosis: a systematic review. JMIR Med Inf 7(3):e10010

Strauss A, Corbin J (1994) Grounded theory methodology. In: Denzin NK, Lincoln YS (eds) Handbook of qualitative research. Sage, New York, pp 273–285

Sturm T, Gerlach JP, Pumplin L, Mesbah N, Peters F, Tauchert C, Buxmann P et al (2021) Coordinating human and machine learning for effective organizational learning. MIS Q 45(3):1581–1602

Tan B, Pan SL, Chen W, Huang L (2020) Organizational sensemaking in ERP implementation: the influence of sensemaking structure. MIS Q 44(4):1773–1809

Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, Shabana W et al (2018) Canadian Association of Radiologists White Paper on artificial intelligence in radiology. Can Assoc Radiol J 69(2):120–135

Vlaar P, van Fenema P, Tiwari V (2008) Cocreating understanding and value in distributed work: how members of onsite and offshore vendor teams give, make, demand, and break sense. MIS Q 32(2):227–255

Weick KE, Sutcliffe KM, Obstfeld D (2005) Organizing and the process of sensemaking. Organ Sci 16(4):409–421

Wiesche M, Jurisch MC, Yetton PW, Krcmar H (2017) Grounded theory methodology in information systems research. MIS Q 41(3):685–701

Yin RK (2009) Case study research: design and methods. Essential guide to qualitative methods in organizational research. Sage, London