To engage or not to engage with AI for critical judgments:

How professionals deal with opacity when using AI for medical diagnosis

Sarah Lebovitz
University of Virginia, McIntire School of Commerce
125 Ruppel Dr., Charlottesville, VA 22903
(sarah.lebovitz@virginia.edu)

Hila Lifshitz-Assaf
New York University, Stern School of Business
44 W 4th St., New York, NY, 10012
(h@nyu.edu)

Natalia Levina
New York University, Stern School of Business
44 W 4th St., New York, NY, 10012
(nlevina@stern.nyu.edu)

Abstract: Artificial intelligence (AI) technologies promise to transform how professionals conduct knowledge work by augmenting their capabilities for making professional judgments. We know little, however, about how human-AI augmentation takes place in practice. Yet gaining this understanding is particularly important when professionals use AI tools to form judgments on critical decisions. We conducted an in-depth field study in a major US hospital where AI tools were used in three departments by diagnostic radiologists making breast cancer, lung cancer, and bone age determinations. The study illustrates the hindering effects of opacity that professionals experienced when using AI tools and explores how these professionals grappled with it in practice. In all three departments, this opacity resulted in professionals experiencing increased uncertainty because AI tool results often diverged from their initial judgment without providing underlying reasoning. Only in one department (of the three), did professionals consistently incorporate AI results into their final judgments, achieving what we call *engaged augmentation*. These professionals invested in *AI interrogation practices* – practices enacted by human experts to relate their own knowledge claims to AI knowledge claims. Professionals in the other two departments did not enact such practices and did not incorporate AI inputs into their final decisions, which we call *un-engaged "augmentation*." Our study unpacks the challenges involved in augmenting professional judgment with powerful, yet opaque, technologies and contributes to literature on AI adoption in knowledge work.

Keywords: artificial intelligence, opacity, explainability, transparency, technology adoption and use, uncertainty, innovation, professional judgment, expertise, decision making, medical diagnosis.

Acknowledgments: We wish to thank the special issue editors and the anonymous reviewers for their invaluable insights throughout the review process. This research benefited from the helpful feedback provided by Beth Bechky and Foster Provost as well as constructive comments from researchers at the NYU Qualitative Research Seminar, NYU Future of Work Seminar, Stanford Changing Nature of Work Workshop, ICIS 2020 AI in Practice PDW, and in the Work in the Age of Intelligent Machines (WAIM)

community. Finally, we thank the individuals at "Urbanside" who graciously allowed us to study their daily work.

INTRODUCTION

Artificial intelligence (AI) technologies are edging closer to human capabilities and are often positioned as a revolutionary resource promising continuous improvements in problem-solving, perception, and reasoning (Rai et al. 2019). These technologies are seen as enablers of a fundamental organizational transformation (Faraj et al. 2018, Kellogg et al. 2019, von Krogh 2018), especially when it comes to professional work (Barley et al. 2017, Erickson et al. 2018). Heated debates are emerging around whether, over time, AI technologies are more likely to "automate" professional work on certain tasks by fully replacing human input, or to "augment" it by keeping human experts in the loop (e.g., Brynjolfsson and Mitchell 2017, Kellogg et al. 2019, Seamans and Furman 2019). Private and public organizations increasingly opt for human-AI augmentation, assuming it will generate value through the synergistic integration of the diverse expertise that AI and experts each offer. In this paper, we study how human-AI augmentation for critical decisions unfolds in practice by closely investigating how professionals use AI tools to form three different medical diagnosis judgments.

Human-AI augmentation is increasingly depicted as "human-AI collaboration" (e.g., Puranam 2021, Raisch and Krakowski 2021, Wilson and Daugherty 2018), emphasizing the need to integrate potentially divergent viewpoints. Drawing on the organizational literature on collaboration, we know that such integration involves transforming knowledge -- a process that requires both understanding the meaning behind others' inputs and being willing to change one's initial position (Carlile 2004, Hardy et al. 2005, Levina 2005, Maguire et al. 2004). It is well-known that achieving effective collaboration in knowledge work is difficult as experts cannot always explain their reasoning due to the tacit nature of knowledge (Polanyi 1958, 1966), and their collaborators may not be willing to listen to unfamiliar viewpoints (Carlile 2004, Levina 2005, Maguire et al. 2004).

The problems of establishing an understanding across diverse bases of expertise and being open to alternative viewpoints are exacerbated in situations when the reasoning behind them is inaccessible.

This is particularly likely to occur when humans face a divergent viewpoint expressed by an AI tool – the so-called "opaque AI" problem. Modern AI tools, such as deep-learning algorithms, often appear as "black boxes" to users since it may be very difficult or even impossible to examine how the algorithm arrived at a particular output (Christin 2020, Diakopoulos 2020, Pasquale 2015). While experiencing opacity and using "black box" technologies (e.g., cars or computers) is ubiquitous (Anthony 2021), problems arise when there is a need to integrate diverse knowledge claims into a single decision that a human expert can stand behind. This is the case for many scenarios of AI use for critical decisions such as in medicine, human resource management, and criminal justice, where opacity associated with AI use is particularly problematic (Christin 2020, Van Den Broek et al. 2021, Waardenburg et al. 2018).

In professional collaboration, human experts integrate diverse knowledge by developing joint practices based on shared interests and common understandings (Bechky 2003a). This enables them to engage in dialogue, at least partially uncovering one another's reasoning in order to arrive at a joint decision. But what would it take for human experts to be able to transform their knowledge based on inputs from black box machines? We set out to explore how experts using AI tools are dealing with opacity and considering whether to alter their initial knowledge claims based on the AI input.

Following a rich tradition of organizational studies investigating technology in work practices (e.g., Barrett et al. 2012, Leonardi and Bailey 2008, Lifshitz-Assaf 2018, Mazmanian et al. 2013, Orlikowski 1992), we conducted an ethnographic field study within a major tertiary hospital in the United States that is using AI technologies for diagnostic radiology. Medical diagnosis in general and diagnostic radiology in particular represent one of the premier examples of professional work that is expected to undergo dramatic transformation as AI technologies continue advancing[1]. We investigate radiologists' use of AI tools for diagnostic processes in three different departments, focusing on their work practices in diagnosing lung cancer, breast cancer, and bone age.

[1] See the Radiological Society of North America's journal *Radiology: Artificial Intelligence* (https://pubs.rsna.org/journal/ai) overview of the state of the field when it comes to AI use and other resources curated by the American College of Radiology (www.acrdsi.org/).

We show how radiologists invested their efforts into reducing uncertainty when forming their diagnosis judgments and how the opacity they experienced when using AI tools initially increased this uncertainty in all three settings. Of the three departments we studied, only in one (when diagnosing lung cancer) were the professionals able to use AI results to enhance their own expertise – the stated goal of the human-AI augmentation. This was a case of what we call *engaged augmentation*, where professionals were regularly integrating the AI knowledge claims with their own. They were able to relate AI results to their initial judgment and reconcile divergent knowledge claims by enacting "AI interrogation practices," which required a significant resource investment on behalf of the professionals who were already highly overextended in their daily work. In the other two departments (when diagnosing breast cancer and bone age), professionals enacted what we call *un-engaged "augmentation"* where they were either regularly ignoring AI's input or accepting it without much reflection. Our study contributes to the nascent understanding of human-AI augmentation practices by unpacking how humans experience and deal with opacity when using AI tools.

BACKGROUND LITERATURE

Augmenting professional expertise with AI

Two scenarios of AI use, either through automation or augmentation, are increasingly debated across academic, practitioner, and policy communities (e.g., Benbya et al. 2021, Brynjolfsson and Mitchell 2017, Cremer and Kasparov 2021, Raisch and Krakowski 2021). In this study, we concentrated on the augmentation scenario, which the literature largely equates with "human in the loop" AI use whereby human experts and AI technologies work together to accomplish a task. The word augmentation is defined as a process of enlargement or making something grander or more superior (Merriam Webster 2021). Indeed, scholars describe human-AI augmentation as an expansion of expertise or knowledge where humans and machines "*combine their complementary strengths*" to "*multiply their capabilities*" (Raisch and Krakowski 2021, p. 6). Through this expansion of expertise, human-AI augmentation is

expected to positively impact organizations through superior performance or improved efficiency (e.g., Brynjolfsson and McAfee 2014; Daugherty and Wilson 2018; Davenport and Kirby 2016).

Embracing the vision of multiplying diverse expertise, many scholars describe human-AI augmentation as humans and machines "collaborating" together (e.g., Boyaci et al. 2020, Gao et al. 2021, Khadpe et al. 2020, Puranam 2021, Wilson and Daugherty 2018). Prior organizational literature on effective collaboration among diverse *human* experts shows how experts learn ways of working together to leverage and combine the complementary capabilities (Hardy et al. 2005, Maguire et al. 2004). Effective collaboration in knowledge work involves transforming and integrating knowledge through a process of relating the knowledge of others to one's own knowledge (Carlile 2004, Levina 2005; Levina & Vaast 2005). This requires collaborators to be willing and able to understand the meaning behind others' input as well as to potentially change one's knowledge claims (Carlile 2004, Levina 2005). A collaboration that effectively integrates divergent knowledge results in individuals not only "adding to" but also "challenging" one another's input, which is distinguished from merely "ignoring" input without reflection (Levina 2005). Extending this literature to AI use, the expectation is that human experts "collaborating" with AI tools are transforming their knowledge by integrating AI results in a way that potentially challenges an expert's initial judgment. Indeed, Raisch and Krakowski assert this expectation when describing augmentation as a tight coupling of human experts and machines influencing one another, wherein "machine outputs are used to challenge humans, and human inputs to challenge machines" (2021, p. 24).

Transforming knowledge is challenging when collaborators are unable to interrogate the other's knowledge claims. Human experts develop collaboration practices based on their shared interests and common understandings that allow them to deliberate each other's knowledge claims (Carlile 2004, Levina 2005, Maguire et al. 2004) despite their inability to fully explicate their reasoning (Polanyi 1958, 1966). While we have been investigating how knowledge workers deal with tacit knowledge over the last three decades of organizational scholarship (e.g., Kogut and Zander 1992), we know relatively little about dealing with the opacity of modern technologies.

Opacity and AI technologies

Issues of opacity, or the antithesis of transparency, associated with organizational adoption of modern technologies has increasingly been a topic of discussion and concern in many research and practitioner communities (e.g., Albu and Flyverbom 2019, Leonardi and Treem 2020, Turco 2016, Zuboff 2015). Opacity refers to the difficulty to understand the reasoning behind a given outcome when such reasoning is obscured or hidden from view (Stohl et al. 2016). While initially, researchers argued that the use of information technology will lead to increased transparency – as more information about activities and decision making was captured digitally and could potentially be accessed and examined by third parties – recent writings have pointed out the fallacy of this thinking (Hansen and Flyverbom 2015, Leonardi and Treem 2020, Stohl et al. 2016). Studying social media platforms as an example, Stohl et al. (2016, p. 125) identify a "transparency paradox," arguing that, while increased use of information technology may increase how *visible* information may be, in certain cases, it may actually reduce *transparency*. This line of argument may be extended to the adoption and use of modern AI tools. Today, such tools are developed with the aim of transforming the glut of "big data" into a digestible piece of highly relevant information –the algorithmic output. Today, these outputs are often presented to users with minimal transparency into how the AI tool generated them. And yet, due to constraints of limited time and bounded rationality, even if all the data and logic underlying an algorithmic output became accessible, transparency may still not be likely (Leonardi and Treem 2020).

The concept of opacity has gained prominence in the context of organizational adoption of AI tools (e.g., Burrell 2016, Christin 2020, Faraj et al. 2018), especially those tools that use deep learning methodologies. These methods often rely on numerous algorithms calculating weighted probabilities that are transferred and transformed through complex multi-layered networks before a given output is generated for users. AI tools using such methods are often referred to as "black boxes" because they may generate unexpected or surprising outputs that end-users and even AI developers are unable to explain or understand (Diakopoulos 2020, Dourish 2016, Pasquale 2015). In the current literature, opacity of AI tools typically describes the lack of explanations provided as to "why a specific decision was made that

are understandable to users, even when they have little technical knowledge" (Glikson and Woolley 2020, p. 631). This focuses on enacted moments of AI use, whereby individuals lack the practical ability to know the reasoning behind *a specific AI result* presented to them, which is distinct from how individuals may lack the ability to evaluate a particular AI tool when examining its technical methodology, training and validation data, and performance measures (Lebovitz et al. 2021).

While the goal of achieving transparency in AI tools seems more necessary than ever – as more and more critical judgments are involving AI tools – the ability to achieve this goal seems more elusive than ever. Scholars, including some computer scientists, are now discussing AI's "fundamental opacity," arguing that transparency may be technically infeasible (e.g., Ananny and Crawford 2016, Xu et al. 2019). Supporters of this view argue that, given the growing complexity of methods and input datasets, "there may be something, in the end, impenetrable about algorithms" (Gillespie 2014, p. 192). Some scholars go so far as to say that achieving transparency in the use of AI is so difficult is that it may be necessary "*to avoid using machine learning algorithms in certain critical domains of application*" (Burrell 2016, p. 9, emphasis added). Not only computer scientists, but scholars from a wide range of fields including law, ethics, political science, information sciences, and management are arguing that using AI for judgments with serious individual or societal consequences may be problematic. This challenge has led to the creation of multi-disciplinary research communities focused on issues of transparency, ethics, and fairness in technology (e.g., Caplan et al. 2018, Crawford et al. 2019). The research in this community broadly covers three areas.

The first area explores how the design of algorithmic models can be more transparent to help address issues of fairness and social justice (e.g., Barocas et al. 2020, Bird et al. 2020, Kaur et al. 2020). For instance, some scholars in this area are focused on developing models that can show how unjust outcomes produced by machine learning models are highly related to bias that exists in the training data. Despite this community's progress using advanced computational methods to improve transparency towards fairness and equality (e.g., Fernández-Loría et al. 2020, Hooker et al. 2019, Samek et al. 2019), most AI tools (and the potential impact of their results on such issues) are still perceived as largely

opaque by their users. This is due in part to the growing computational complexity of deep learning models and the "curse of dimensionality" when attempting to assert what features from massive sets of input data are yielding specific predictions (Domingos 2015).

The second area of research explores the relationship between algorithmic transparency and professional accountability (Diakopoulos 2020, Pasquale 2015). This work is based on the reasoning that a system can be better governed if its inner workings are more transparent and known (Ananny and Crawford 2016). This is critical since introducing AI tools into a professional work setting may transform existing distributions of responsibility and accountability without providing the ability to view or understand the underlying logic (e.g., Ananny and Crawford 2016, Caplan et al. 2018, Scott and Orlikowski 2012). Related questions are also being raised about the impact of opacity on new forms of algorithmic management and control, as workers are often unaware of how algorithms are directing and evaluating their work (Kellogg et al. 2019, Watkins 2020).

The third area of research focuses on classifying and characterizing the types and sources of transparency and opacity associated with AI systems. Some work in this area has focused on distinguishing, for example, between the transparency of a system's training datasets from transparency about the specific features and weights that led an algorithm to a given outcome (e.g., Diakopoulos 2020, von Krogh 2018). Another area within this topic has investigated the reasons behind opacity of AI tools, such as intentional organizational or managerial secrecy, technical complexity of the tools, and structural factors that pre-existed the AI system, among other reasons (Burrell 2016, Christin 2020).

Today, despite the enduring challenges of opacity, AI tools are increasingly being implemented in contexts where professionals are expected to integrate their own knowledge with AI results when forming judgments in critical contexts (Nunn 2018, Razorthink Inc. 2019). Prior research has shown knowledge workers attempting to examine the underlying logic as they encounter new technologies, such as digital simulation technology in manufacturing (Bailey et al. 2012) and engineering (Dodgson et al. 2007). However, in modern contexts of human-AI augmentation, professionals are expected to "collaborate" and transform knowledge without the practical ability to examine or evaluate AI knowledge claims. Thus, our

study focuses on the question: how do professionals experience and deal with opacity when using AI tools to form critical judgments?

Investigating opacity of AI-in-use through sociomaterial practices of knowledge work

To investigate this question, we focus theoretically on the sociomaterial practices of knowledge work that AI tools are involved in. We adopt a relational ontology that assumes the entangled nature of actors and materials and foregrounds the performativity of practices (Barad 2003, Suchman 2007). This perspective emphasizes the way in which technologies and actors are inseparable and continually (re)produce one another through practices situated within particular social and historical contexts (Leonardi 2011, Orlikowski 2007, Orlikowski and Scott 2008, Suchman 2007). This lens has been used to uncover important insights when studying organizational uses and impacts of other technologies such as enterprise integration platforms (Wagner et al. 2010, 2011), social media tools (Scott and Orlikowski 2014), online community platforms (Barrett et al. 2016), and robotic tools (Barrett et al. 2012, Beane and Orlikowski 2015). We follow Suchman's (2007, p. 1) argument to shift from "categorical debates", in our case, around AI and opacity, to "empirical investigations of concrete practices" in which individuals and technologies act together.

Adopting this view means focusing on the generative materiality of technical infrastructures and treating the technologies-in-use (AI and otherwise) as part of the sociomaterial configuration (Barrett et al. 2016, Mazmanian et al. 2014, Scott and Orlikowski 2014). In particular, focusing on situated configurations emphasizes that individuals' understandings about a given technology vary across local meaning systems (Leonardi and Barley 2010, Mol 2003, Pinch and Bijker 1987). This means leaving opacity to be realized in practice "depending on the actor's situatedness" (Haraway 1988). Therefore, instead of conceptualizing opacity as inherent or fixed features of AI tools, we view opacity as something produced and enacted through practices situated in specific organizational configurations (Leonardi 2011, Orlikowski 2000). Using this lens, we set out to examine how opacity of AI-in-use is experienced and dealt with when professionals use AI when forming judgments.

METHODS

Research setting

We conducted an in-depth field study within three different departments in a large diagnostic radiology organization at Urbanside, a teaching hospital in a major US city. Diagnostic radiology is a specialized medical field in which medical imaging is analyzed to diagnose and treat diseases, and it has been at the forefront of adopting cutting-edge technologies (AI and non-AI) for decades (e.g., Barley 1986). Recently, a great debate has been unfolding as to the impact of AI tools on professionals in this field and how AI may entirely replace the radiology profession (Grady 2019, Mukherjee 2017, Recht and Bryan 2017). We designed our study, following the tradition of field studies of technologies, work, and organizations (Barley 1990, Bechky 2019, Lifshitz-Assaf 2018, Orlikowski 2000), to investigate three radiology departments within the same organization and enabled us to deepen our investigation of professionals' work with AI tools.

Data collection

Starting in late 2018, we immersed ourselves in the field of diagnostic radiology, attending professional conferences, symposia, and vendor events, to understand the opportunities and challenges on the professional field's horizon. Ethnographic field work began in January of 2019 and studied 40 radiologists (licensed doctors or senior fellows offered positions upon completing their fellowship) across three departments actively using AI tools: breast imaging, chest imaging, and pediatric imaging.

***Observation*.** The primary source of data for this study is ten months (over 500 hours) of ethnographic observation (Van Maanen 1988). We documented over 1,000 cases of radiologists forming diagnoses in detailed written observational notes, which were transcribed and supplemented upon leaving Urbanside facilities each day. Because Urbanside radiologists trained medical students and residents, we often captured radiologists verbally articulating their diagnostic reasoning, drawing on past experiences and research, describing common errors and strategies to avoid them, and so forth. Radiologists often quizzed trainees about important diagnostic practices and philosophies (e.g., "What might hypo-inflation indicate in a newborn?" or "What might indicate stroke on MRI?") and then offering their own thoughts.

During periods of observation, we paid close attention to the technologies-in-use, capturing the role of the tools in the diagnostic process, the results they produced, what meanings emerged around the tools, and so forth. Over the course of our field work, we observed diagnostic cases involving and not involving AI tools. Observing cases not involving AI tools strengthened our understanding of radiologists' analytical practices. Even for diagnosis scenarios typically involving AI tools, we also observed cases of radiologists not using the tools, such as during technical outages or when working for satellite locations with different technical infrastructures.

Interviews. Observational data was enriched through 33 semi-structured interviews (Spradley, 1979). Twenty-one informal interviews took place as radiologists conducted their work or during short breaks, covering questions about unclear aspects of diagnoses for recent patient cases, interactions with their colleagues or patients, or specific moments of using or not using various technologies. Twelve formal interviews allowed us to deepen our understanding of what it means to be a radiologist, how they go about their diagnostic work, their perceptions of various technologies, and so forth. All formal interviews and some informal interviews were recorded (with informants' consent) and transcribed.

Documentation and artifacts. Finally, we collected documentation and artifact data which served multiple purposes in our study. First, we captured artifacts produced and used by radiologists in their daily work, including medical notes and photographs or drawings of medical images they were referencing. These materials supplemented observational notes and strengthened our analysis when reconstructing their diagnosis process. Next, we collected technical research papers, regulatory filings, and vendor documentation to study the three focal AI tools and the nature of their outputs. In the US, once regulators approve a clinical AI system, it can no longer change (or "actively learn"). Vendors can request additional approval for updated software versions, which can then be deployed in clinical settings. Thus, we observed the use of unchanging technologies throughout our study.

Data analysis

In keeping with the principles of grounded theory development, we engaged in iterative rounds of data analysis during and throughout our data collection (Charmaz 2014, Glaser and Strauss 1967). In the early

stages, we conducted open coding to capture a broad range of emerging themes. Within the first few months, the prominence of radiologists expressing doubt, asking questions, double-checking, and conducting deep analytical practices was striking in the data. We were also struck by the frequency of questions and confusion surrounding the AI results radiologists viewed. We, therefore, conducted targeted rounds of data collection and analysis to deepen our understanding of these themes.

While all radiologists appeared to be "using" the AI tools (clicking to display its results after forming their initial judgment), we noticed different patterns in the degree that AI results were influencing radiologists' final judgments (e.g., "pausing to consider AI results," "updating original diagnosis," "quickly disregarding AI results"). In all three departments, the AI results and the radiologists' opinions often diverged, and confusion and frustration often followed. Deeper analysis led us to relate their frustration to the lack of understanding why a given AI result was produced (e.g., "questioning what the AI is looking at," "guessing factors behind AI output"). When we investigated the three AI tools and the nature of their output, we found many similarities: each reported high performance metrics, used neural network classification methods, and offered no explanation of its results to users. And yet, despite similarities in the tools and radiologists' consistent frustration, only radiologists diagnosing lung cancer were regularly incorporating AI results, while, the other radiologists mostly ignored the tools' results.

Next, we set out to understand what was behind these divergent patterns. We mapped step-by-step how radiologists formed each different type of diagnosis and analyzed their process along multiple dimensions, such as what aspects of the diagnosis prompted doubt, how evidence was analyzed, perceptions of the AI tool and its results, and so forth. We studied radiologists' similarities and differences among the diagnostic settings and their analytical practices and saw noteworthy differences in the materialities of the imaging technologies-in-use (CT scans, mammography, and X-ray) and the breadth and depth of analysis that was afforded. Iterating with the literature on professional adoption of technology led us to analyze how senior and junior radiologists used the tools similarly and how all radiologists held similar attitudes about AI adoption. Further analysis led us to focus on a key difference in how radiologists integrated the AI result (or not) using what call "AI interrogation practices". We

continued to sharpen our analysis by consulting literatures on epistemic uncertainty and opacity, which further enhanced our formal theory development, which we describe in the following section.

FINDINGS

Diagnosing patients is a critical process that requires the extensive expertise, training, and judgment of diagnostic radiology professionals. Radiologists develop deep expertise in diagnosis through at least six years of intense, immersive education after medical school. In their daily work, they strive to provide the best possible care to their patients and take their role in patients' health outcomes very seriously. They work under resource constraints and time pressure, as healthcare facilities respond to intense pressures to increase patient volumes and reduce costs. In recent years, powerful diagnostic AI tools have captured the attention of radiologists and healthcare leadership. We present below how radiologists in three departments at Urbanside worked with AI tools to provide three critical types of medical diagnoses.

I. Producing lung cancer diagnoses using AI tools

Diagnosing lung cancer was a key focus of Urbanside radiologists specializing in chest imaging. Like others across the field, these radiologists were committed to producing the most accurate diagnoses possible and positively impacting patients' treatment and health outcomes. As in other Urbanside departments, they faced high workload demands and felt strong pressure to work quickly. At the same time, they provided thorough analyses requiring intense concentration and careful deliberation. When diagnosing lung cancer, radiologists faced the challenging task of identifying difficult-to-detect lung "nodules" and characterizing their likelihood of malignancy. Radiologists were deeply aware of the significant consequences of their diagnoses, both the cost of falsely diagnosing a healthy patient and the cost of missing signs of cancer, and worked with high diligence.

Forming critical judgments (without AI): Experiencing high uncertainty

While forming a lung cancer diagnosis, radiologists experienced three main sources of uncertainty when reviewing their primary source of evidence, CT imaging: multiple series of high-resolution images (in 5mm and 1mm "slices") that digitally reconstructed 3D cross-sections of a patient's upper body and

supported numerous settings and projections (e.g., from side or overhead views, varying degrees of contrast).

First, they experienced great uncertainty while discerning "lung nodules" from the healthy lung tissue. This involved searching for small white-appearing circles within the varying shades of white to dark grey lung tissue on the CT images. However, hundreds of small white circular areas may be visible on a given CT that represented normal tissue or bone (see Appendix 1), and radiologists often wavered considerably while deciding whether a particular area was a nodule or not. One afternoon, a physician called Dr. E's phone, requesting her opinion about a potential nodule on her patient's CT. After several moments of searching and deliberating over the phone, Dr. E asked the physician, "Do you mind if I look more closely and figure it out and call you back?" Hanging up, she leaned closer to the monitors and continued her analysis before finally returning the call: "It's very low density. It's looking almost fat-like [which appears more grey than typical a nodule]. But it actually does look like a nodule. Sometimes I'm like, 'Am I going crazy?'"

The second source of uncertainty emerged from radiologists' task of identifying each and every nodule in the patient's lung tissue. Very frequently, they expressed concern about the possibility of missing a nodule, fearful of making consequential errors of omission: "I don't see anything major jumping out. Hopefully I'm not missing anything" (Dr. Y). This struggle was related to the CT imaging not always clearly capturing every region of a patient's lung tissue where nodules may be positioned, as in the following case of Dr. E deliberating aloud: "Am I hallucinating a nodule?…I think it's there, but it's hard to see. It's in a bad location…It's behind two ribs, so it's impossible to get a good look there." Dr. J explained how "there's all this lung tucked in front and behind right there that you just don't see [on CT imaging]." Radiologists' difficulty examining these "impossible" areas of lung tissue using CT imaging raised their uncertainty. In fact, they often concluded that a seemingly-nodule free CT scan was not definitive: "If you don't see the nodule on one image, that doesn't mean it's not there…A lot of missed cancers, like ten percent, were seen only from one view and not the other" (Dr. J). The CT, like other imaging technologies, may also be difficult to analyze when patients shift or fail to inhale deeply

during the scan, as in the case of Dr. S struggling to discern a particularly blurry CT image: “It’s hard to tell because it’s such a crappy study. He did not take a deep breath, did he?”

Radiologists worked to address these first two sources of uncertainty by investing in various analytical practices during the “nodule search”. They methodically combed through the CT images numerous times, starting with the less granular set of 5mm images and then the more granular 1mm set, as Dr. J explained to medical students observing her work: “There is so much volume of data on the images to deal with…We scroll faster at first. It’s good to get a general overview first, then we go to the smaller ones for deeper investigation.” Then, their focus turned to further evaluating each potential nodule they identified, scrolling slowly through the neighboring slices to assess if it appeared to “flow” in a continuous path (indicating normal blood vessels) or disappear abruptly (indicating a nodule): “You have to follow the vessels. If it’s something you’re able to follow, then it’s probably just a vessel you’re catching, not a nodule.” They increased their confidence using a technique called “windowing”, or assessing the different properties of the tissue by adjusting the settings of the CT image or changing its grey-scale contrast: “Oh, I think it’s a vessel. Yeah, I don’t think it’s a nodule. Ah, yeah, I’m pretty sure. Windowing really helps” (Dr. E). As a final measure to address lingering concern or confusion, they may request additional imaging, as Dr. J explained to on-looking medical students: “This area looks ill-defined. So, somebody could call it a nodule. We try to make a firm guess…but sometimes we call for follow-up imaging because we really can’t decide.”

Finally, radiologists faced a third, and relatively less acute, source of uncertainty during the task of characterizing each nodule’s likelihood of malignancy. Radiologists applied fairly explicit criteria and standards to each nodule they had identified: “Almost everyone has nodules, but some of them can be cancer...You go through each nodule and make sure it’s solid. Then with the prior [CT images] that you’re comparing to, you actually look at each nodule and visually make sure they look the same” [Dr. Y]. They first gauged the patient’s overall risk level by reviewing their medical details (e.g., clinical symptoms, age, history of illness). Next, they scrolled through the CT images several times to explore the nodule. They used digital tools to precisely measure its dimensions and noted whether it was larger than

the 5mm standard associated with malignancy. They analyzed prior CT imaging (if available) looking for changes or stability in the nodule's appearance or size over time. When Dr. J noted that a 3mm nodule was present on a CT scan from five years earlier where it also measured 3mm, she felt highly certain in characterizing the nodule as benign: "There it is [in the CT from 2014]! So it's there [not new]. Oh, that's stable [noting the consistent 3mm measurement]. There it is. Okay, now I'm good."

Experiencing opacity of AI-in-use (and increasing uncertainty)

After completing their initial analysis, radiologists then viewed the results of an AI tool implemented to aid their lung cancer diagnosis. The Urbanside chest imaging department purchased an AI tool several years prior, which we refer to as the "CT AI tool", as an add-on to the CT digital imaging technology, from a leading healthcare technology provider. Over the years, the tool was updated numerous times to improve its technical sophistication and performance. At the time of this study's observation, the tool performed imaging processing, segmentation, and classification tasks utilizing artificial neural networks that were trained and validated using large datasets of long-term radiological outcomes. Published research showed these AI tools' ability to identify and classify nodules was similar to radiologists' cancer detection rates. Following regulatory guidelines, the CT AI tool was deployed as an "aid" to radiologists, designated to be used *after* the radiologist first formed his or her independent judgment.

Clicking an icon on the digital workstation, instantaneously, the display jumped to the first AI result, a circle annotation placed on a precise location of the CT image. In the intermittent cases where the AI result and the radiologists' judgment converged that we observed, they quickly moved on to complete the final report. Radiologists expressed delight and relief when the AI results confirmed their previously uncertain assessment that no nodules were present: "This time, [CT AI] found nothing. Any time that happens, it puts a big smile on my face" (Dr. F). They experienced a boost in confidence and certainty after viewing the AI results, as Dr. W expressed, "If I don't see any lung nodules, and [CT AI] doesn't see any lung nodules, then okay, we're good! I now feel very comfortable saying there's no lung nodules."

However, in the majority of cases we observed, the AI tool's results presented a divergent view from the radiologist's initial view. Regularly, the CT AI tool did not mark a nodule the radiologist had

identified. Even more frequently, the tool flagged additional areas that the radiologist had not identified. Radiologists began experiencing opacity, as they were unable to understand these divergent AI results. They questioned what features of underlying lung tissue were relevant to the tool's decision: "How does [the AI tool] know that *this* is a nodule, but *this* isn't?" (Dr. V). Radiologists were deeply committed to providing judgments with maximum certainty, but they expressed difficulty feeling certain given the opacity they experienced when considering divergent AI results: "I just don't know of any radiologist who's not looking closely at the case because they have AI. Because at the end of the day, you're still responsible. How can you trust the machine that much?" (Dr. E).

Dealing with opacity of AI-in-use: Enacting AI interrogation practices and incorporating AI results

On the surface, it may seem that using the AI tools (and experiencing opacity) increased the overall uncertainty these radiologists experienced; however, in fact, using the AI tool resulted in radiologists experiencing *less uncertainty* making their final judgments. They achieved this by using "AI interrogation practices", or practices that human experts enact to relate their own knowledge claims to the AI knowledge claims. For these radiologists, enacting AI interrogation practices involved building an understanding of the AI result and then reconciling the divergent viewpoints. They examined the suspected area in question, zooming in on that region of the CT image and scrolling forward and backward to assess the tissue surrounding the AI-marked region. They changed the contrast settings on the CT to analyze the area's size, shape, and density and reviewed prior CT images to understand how those features may have changed over time. They were examining and probing the AI results in order to understand them and ultimately integrate them with their own viewpoint.

Enacting AI interrogation practices led radiologists to consistently integrate the AI results into their final judgments. Radiologists regularly updated their initial opinion after interrogating the AI results, either through synthesizing the divergent opinions into a new insight or through reflectively agreeing with the AI result, as in the following case. After completing his initial analysis, Dr. T was puzzled by three AI results suggesting nodules he had not initially flagged. He began interrogating each area marked by the AI tool, analyzing the CT imaging to try to understand the AI result and how it related to his own view. He

decided to overrule one AI result and expand his original opinion to include the two new additional ones. Even when radiologists decided to overrule the AI results, they experienced higher confidence reporting that final diagnosis. This was the case after Dr. F swiftly interrogated two unexpected AI-marked areas and related them to his own analysis: "This is what [CT AI] picked up: there and there. It's just normal stuff, parts of the bones protruding from the chest which sometimes looks like it could be a nodule."

Enacting AI interrogation practices required radiologists to invest additional time and analysis. They were willing to make that investment time and time again, which reflected their positive views of the AI results' value, as expressed by Dr F.: "I know my limitations and I know this [CT AI] is going to help them [nodules] stand out a little better. It's worth the extra time in my mind." They viewed the AI results as distinct and complementary to their own capabilities and expressed strong positive opinions about the tool's value in their work. This was vividly expressed by Dr. W., a senior radiologist who moved from another hospital where they did not have a CT AI tool: "I actually think [CT AI] is mission-critical. For me to read cases, I absolutely love having the [CT AI]. I used to not have it in my prior place [hospital]. I thought it was the worst thing ever. And then when I came here, I was amazed."

Indeed, the practice of interrogating and integrating the AI results had become a critical step in how these experts formed their final judgments. This was reflected in Dr. V's response one afternoon when the AI results were unexpectedly unable to load for a CT she was assessing. She instant-messaged the CT technician, requesting the AI results for that study, and followed up with a phone call when the technician did not respond. She minimized that CT and began analyzing another case while she waited; a few minutes later, she learned of technical issues disrupting the AI services. Flustered, she returned to the minimized case, scrolled through the CT several more times, and reluctantly wrote the diagnosis report without AI input: "Once in a while, it definitely picks up things that you looked at yourself and you totally ignored, that you just couldn't see. Knowing that every now and then it picks up something real makes you always want to go back to it."

II. Producing breast cancer diagnoses using AI tools

As breast cancer is prevalent and highly dangerous, diagnosing it at the earliest and most treatable stage was a great priority for radiologists specializing in breast imaging. On a typical day, each Urbanside breast radiologist evaluated over 100 patients– a highly demanding workload – and was providing life-or-death judgments in every case: "We have to give our full attention to make the right call, but we have so much volume we're supposed to get through. It's a conflicting thing" (Dr. Q). On average, they spent less than three minutes evaluating a case, an amount of time that did not allow extensive deliberations. The consequences of making these evaluations were extremely high, as patients were either informed they were not currently at risk or recommended to undergo additional testing, biopsy, or treatment, which resulted in patients bearing significant physical, emotional, and financial costs.

Forming critical judgments (without AI): Experiencing high uncertainty

While making critical judgments about breast cancer, radiologists experienced two main sources of uncertainty. First, like radiologists conducting the lung nodule search, breast radiologists wrestled with identifying abnormal areas within the complex breast tissue anatomy. The main source of evidence is mammography imaging: digital X-ray imaging that provided four two-dimensional images and four three-dimensional images (side and overhead view of each breast). For certain patient scenarios, targeted ultrasound imaging was also used.

Breast radiologists worked to detect every potential abnormality in the patient's breast imaging and knew that overlooking a single abnormality carried extremely high consequences. On mammography, abnormalities typically appear as small bright white patches amidst normal tissue ranging from white to dark grey (see Appendix 2). Because of the subtle differences in tissue appearance, and the difficulty of interpreting mammogram imaging, radiologists frequently expressed concern about missing critical findings. Dr. C explained, "The [abnormalities] we worry about are really faint and tiny ones: those are signs of early cancer. They're the ones you can barely see…A[n abnormality] is going to be really really masked....you just can't see the cancer.…It's like looking for a snowball in a snowstorm." In some cases, radiologists requested additional imaging to be more sure, especially when the mammogram did not

capture areas of the patient's body (often near the armpit): "If I could see clearly in *this* area [pointing just outside the border of the image], I wouldn't be so concerned" (Dr. L).

To increase their confidence that they identified all abnormalities, radiologists used careful analytical practices. They combed over mammogram images, zooming in closely on each region, and scrolling through each three-dimensional view multiple times. They were searching for unusual patterns in the tissue that may indicate "masses, calcifications, skin thickening, changes to the tissue, axillary lymph nodes, or distortion" (Dr. K). They examined asymmetries in the appearance of the left and right breast tissue, as Dr. P described, referring to images on her screen: "This is an area that caught my eye. This is the right side, and this is the left. The right looks obviously different than the left. This is one of the things that our eyes are trained to look for." Using careful systematic analysis that provided diverse views and evidence helped ward off radiologists' uncertainty, as Dr. G explained: "I zoom in even more, so I'm going to see even the tiniest finding. I zoom in, like, a lot….until I'm pretty sure I see all of them."

The second, and more intense, source of uncertainty was characterizing each abnormality's likelihood of being malignant or benign. Making this distinction for breast cancer diagnosis was challenging. Radiologists described breast cancer as a complex disease that may develop in unexpected ways that often varied from patient to patient. Breast tissue anatomy is complex, and often, malignant breast tissue may closely resemble healthy tissue on mammography. Numerous pieces of evidence needed to be analyzed and synthesized: the size, shape, edges, and density of the abnormality on mammography, ultrasound, and MRI (if available), the degree of change across prior imaging, a patient's genetic makeup, prior history of disease, and lifestyle choices, as well as clinical symptoms and physical examinations.

Occasionally, the evidence would overwhelmingly support a benign judgment, as in this case, "Those calcifications are really big and chunky, so once I look at it closely, then I can immediately ignore it [because it is likely benign]" (Dr. C). More frequently, however, some factors would suggest benign while others did not, and radiologists struggled to reduce the acute uncertainty: "If I know something is fine or I know something is bad, then I know right away. But there are really a lot of cases I waffle on" (Dr. B). This is illustrated by the following case. During her analysis, Dr. C noted a small gray oval

abnormality on a patient's mammogram. Through her analysis, she noted the oval's small size and sharply defined edges, that the area appeared stable for several years, and that the medical history did not suggest increased risk (all suggesting benign). And yet, she felt uncertain and exhaled deeply in frustration before ultimately recommending the patient undergo additional testing: "It's probably normal tissue, but it looks *so oval*. I've gotta call her back [for additional testing]. *I just can't ignore that spot*."

They expressed deeper anguish and deliberation when judging the malignancy of abnormalities than when searching for them, as Dr. Z explained: "Deciding what to *do* with an abnormal finding [deciding malignant versus benign] – as opposed to detecting a finding in the mammogram – that takes much more discerning." Colleagues often disagreed about an area's likelihood of malignancy, especially since the mammogram imaging and its various features were open to multiple interpretations. Even after completing their full analysis, radiologists often second-guessed their final judgment, as portrayed by Dr. G's continued wavering: "Do I do a follow up or do I just return to routine screening? That's really the difference between being a cyst [benign] and something being a solid mass [malignant]. And *we can't always tell the difference*."

To build certainty in this judgment, radiologists used a variety of analytical practices. They zoomed in on the mammogram to examine the appearance of the abnormality and its density, size, shape, and edge clarity. They gauge whether the abnormality had changed or remained stable across prior years' mammograms. They also studied the patient's health records (e.g., physical symptoms, personal and family history, pathology and surgical records) to gauge the patient's overall risk level and inform their emerging judgment. In one case, Dr. L decided to recommend a biopsy after considering a patient's elevated risk factors, despite the area's otherwise benign appearance: "It's not overtly suspicious: it's fairly circumscribed and it's not very oval [both suggesting benign diagnosis]. But this patient is here because she just found out from a genetic risk screen that she is at increased risk for breast cancer."

Experiencing opacity of AI-in-use (and increasing uncertainty)

After forming their initial judgment, radiologists then reviewed the results of an AI tool implemented to aid their diagnosis process. Several years ago, Urbanside purchased an AI tool, which we call the

"Mammo AI tool," as an add-on product to the mammography software from the imaging technology vendor, one of the leading US healthcare technology providers. Since its implementation at Urbanside, the vendor provided numerous updates improving the tool. During this study's observations, the Mammo AI tool performed imaging processing, segmentation, and classification tasks utilizing artificial neural networks trained and validated using largescale datasets with long-term radiological outcomes. Published research reported that the tool could identify malignancies at similar rates as trained radiologists and showed some indication of increasing radiologists' overall cancer detection rates[2]. Following regulatory guidelines, the tools were deployed as an "aid" to radiologists, who were required to only view AI results after forming their independent evaluation. The tool was designed so that a single mouse click displayed the AI tool results: a series of shapes[3] marking the specific location on the mammogram that was classified as malignant, with no further information (Appendix 3).

Clicking the designated button, the AI results appeared on the mammogram image, which the radiologist compared to her initial judgment. In the infrequent cases we observed where the opinions converged, they swiftly proceeded to the final diagnostic report. However, in the large majority of cases we observed, the AI results and the radiologists' judgment diverged. On occasion, the AI tool did not flag an area that was initially judged as abnormal, and far more frequently, the AI tool flagged additional areas that the radiologist had not.

Radiologists experienced opacity as they encountered the AI tool's unexplained results. They were unable to see what aspects of that tissue were causing the AI tool to produce a given result: "I don't know why they marked these calcifications, what about all these other calcifications (that the tool did not mark)? They all look identical to me" (Dr. C). They expressed frustration in their inability to understand the divergent AI results: "What is it telling me to look at? At this tissue? It looks just like the tissue over

[2] This research prompted the US government in 2003 to mandate that insurance providers must reimburse the use of AI tools for breast cancer screening, leading to wide purchasing of such tools across US breast imaging centers.
[3] Three shapes were used to indicate the type of classification the tool generated: star indicated "mass," triangle indicated "calcification," and plus sign indicated cooccurrence of mass and calcification.

here, which is perfectly normal…I have no idea what it's thinking" (Dr. K). Radiologists had no practical means of knowing the underlying reasoning of a given AI result and experienced the opacity of AI-in-use, as Dr. H explained, "[The AI tool] just points an area out and leaves you to figure it out. It's like it's saying, 'This is weird; what do you want to do with it?'"

Dealing with AI opacity: Not enacting AI interrogation practices and not incorporating AI results

Like chest radiologists' use of the CT AI tool, in this department, breast radiologists using the Mammo AI tool experienced opacity and a surge in their level of uncertainty. However, in this department, radiologists did not enact AI interrogation practices and ultimately did not regularly incorporate AI results into their final judgments. Instead, when faced with divergent opinions, the radiologists tended to review the image underlying the AI result in a perfunctory way before ignoring it, or "blowing it off", as Dr. G described, "I blow so many things [AI results] off. Like if there's normal scarring or stable calcs [benign tissue], it's [AI tool] going to pick up *everything*." They quickly dismissed AI-marked areas that they previously deemed normal without deeper inspection, writing them off as "false positives": "I already knew that stuff it marked didn't matter. I saw the mass was there a couple of years ago [in prior imaging]" (Dr. Z). It was also common for them to ignore AI results when the AI tool did not flag an area they initially considered abnormal: "If there's something that's concerning to you, based on your initial interpretation, that the [AI tool] is saying, 'Oh, this looks normal,' you couldn't use that information and say, 'We're *not* going to biopsy it'" (Dr. I).

Radiologists already faced extreme uncertainty and intense time pressure, which suddenly multiplied when they had to reconcile the (frequently) divergent opinions of the Mammo AI tool: "So many different factors are standing out to you all at once and giving you conflicting information, *and then* there's the result from the software [Mammo AI]" (Dr. L). They expressed strong opinions that the Mammo AI results did not add value to their process, based on years of repeatedly spending valuable time reviewing divergent and unexplained AI results: "It isn't helping anybody. It's actually just another step for me to do" (Dr. K). Radiologists expressed negative views of having to tediously check, and ultimately "blow off", AI results for every patient's case, especially given the high time pressure they faced: "It's

not worth my time to evaluate it" (Dr. L). Only under specific conditions (when analyzing highly dense breast tissue), did some radiologists comment on the potentially complementary nature of the tool's results: "Calcifications[4] can be really little and sometimes hard to see. It [Mammo AI tool] sees those calcifications better than I do. But it also sees all kinds of calcifications that are *neither here nor there*" (Dr. B). And yet, in the same breath, Dr. B conveyed her view (shared by most of her colleagues) that the AI results were often useless when making her final judgments (they were "neither here nor there").

In the end, due to the lack of full feedback on patients' health over time, it is unclear whether radiologists' decisions to not incorporate AI results led to more effective treatment or not. It is possible that for some cases, had an AI result been incorporated, additional patient testing may have been avoided. For instance, Dr. L was examining new images for a patient who had been recommended for additional imaging by Dr. L's colleague the week before. Dr. L opened the patient's original mammogram (from the prior week) and reviewed the AI output, which had not flagged the area that prompted her colleague's concern: "[Mammo AI] didn't mark *anything* on this one [the prior week's image]. It didn't even mark the lesion that caught the radiologist's attention!" (Dr. L). Interestingly, after Dr. L reviewed the patient's new images, she recorded her opinion that the area was benign. This pattern was not uncommon: radiologists often recorded benign judgments after reviewing additional imaging. In this case, the original AI result was consistent with the radiologist's ultimate benign diagnosis, however, its accuracy is unclear without long-term patient health outcomes.

III. Producing bone age diagnoses using AI tools

"Bone age" evaluation involves radiologists specializing in pediatric imaging to assess the skeletal maturity of children experiencing delays in growth or puberty. This important diagnosis factors into considering whether to treat the child with daily growth hormone injections for a period of time. This

[4] Calcifications are tiny flecks of calcium that can sometimes indicate early signs of cancer. They are usually unable to be felt by a physical examination. Large calcifications are not usually associated with cancer. Clusters of small calcifications indicate extra breast cell activity, which can be related to early cancer development but may also be related to normal breast cell activity.

diagnosis involves comparing a child's bone development to established pediatric standards to determine whether it falls within a "normal" or "abnormal" range for the patient's age. A pediatric radiologist at Urbanside may perform seven or eight bone age evaluations on a given day, among the variety of 40-50 other diagnoses they provide (e.g., evaluating lung disease on CT scans, gastrointestinal issues on Ultrasound, or scoliosis on X-Ray). Like in the previous departments, these radiologists faced acute pressure to work quickly and provide high-quality, time-sensitive assessments to physician teams caring for young patients and their concerned parents.

Forming critical judgments (without AI): Experiencing lower uncertainty

Unlike the previous two specialties, pediatric radiologists viewed this evaluation as a straightforward comparison task and did not experience particularly high uncertainty, or as Dr. O described, "I don't think it's a very sophisticated thing." After first quickly noting the patient's age and gender, they reviewed the sole source of evidence for this evaluation: a single digital X-ray representing the patient's hand, fingers, and wrist (see Appendix 4). They studied the size, shape, and appearance of specific bones visible on the X-ray and drew on their knowledge of how certain parts of the hand develop differently over time: "I use the phalanges [fingers] as the gold standard, but there's also carpal bones [wrist] and the radius ulna [forearm]. But they're more variable, so I don't look at that as much" (Dr. R). Dr. D explained how she considers and weighs multiple bone areas, to build a more certain judgment: "I give more credence to distal bones [closer to fingertips], although endocrinologists like the proximal [lower fingers or wrist], which is probably more representative of overall height growth…If there's variation, or if there's discordance between different bones, I mentally give more weight to some than others."

Then, they compared the patient's bone development to the curated set of X-ray images in the textbook of standards used across pediatric radiology. A single X-ray image was used to depict a child's expected bone development at each one-year increment. Radiologists compared the appearance of the patient's hand X-ray to the standard imagines in the book, searching for the closest match: "I'm looking at the different shapes and seeing these are bigger than seven years" (Dr. D). In the following assessment, Dr. N went back and forth between the 18- and 19-year standards, noting slight differences in the bone

development: “You see here, the bones are all fissured [pointing to patient’s X-ray]. And here [in the 18-year standard image], there’s still a tiny physis[5].” A faint white line (the “tiny physis”) ran horizontally between the knuckle and fingertip in the standard 18-year-old image, but no horizontal line appeared on the patient’s X-ray (it was “all fissured”, or no gap between the bones). Dr. N interpreted this to mean the patient’s bone age was greater than 18 and thus reported his judgment of 19 years.

Lastly, radiologists performed a calculation of the “normal” range of bone ages using a data table of standard deviations for each consecutive age printed in the textbook and reported whether their judgment of the patient’s bone age fell within or outside that range.

Experiencing opacity of AI-in-use (and increasing uncertainty)

After forming their initial judgment, the radiologist then viewed the result of the AI tool. In 2018, the Urbanside pediatric department implemented a cutting-edge tool, which we refer to as the “X-ray AI tool”, to aid in bone age diagnosis. Citing the fairly straightforward comparison or “pattern recognition” nature of this task, Urbanside pediatric radiologists expressed high enthusiasm for using the X-ray AI tool, as Dr. N explained: “I think [the AI tool] can be very useful…You have to look very finely and carefully at a bunch of different images. It’s visually overwhelming. But I think it’s something a computer is really good at...It’s just pattern recognition.” The tool was developed by a reputable research institution and used deep learning methods at the forefront of diagnostic AI development at the time, involving multiple stages of convolutional neural networks performing image processing, segmentation, and classification tasks. Published studies reported the tool’s results matched the “normal” vs “abnormal” judgments of pediatric radiologists in over 95% of test cases. Urbanside radiologists eagerly agreed to participate in a multi-institution effort to further study the tool in settings of clinical use.

Once implemented, every bone age evaluation was automatically processed and analyzed by the X-ray AI tool before entering the radiology work queue. Upon opening a bone age case, the digital X-ray

[5] A physis is a growth plate located between bones. Over time, the physis becomes thinner until eventually disappearing as one nears full growth.

displayed on the center monitor and the diagnostic report software loaded on the side monitor. The X-ray AI tool automatically populated the diagnostic report with the AI result, a specific bone age measurement, and its corresponding "normal" or "abnormal" evaluation. Like in the previous two cases, radiologists first formed their initial opinion and then viewed the AI result and decided how to use it.

Viewing the AI results, all of a sudden, radiologists experienced a new surge of uncertainty, rooted in their inability to understand or explain the AI result. In about a third of the cases, the AI tool's bone age roughly converged with their initial judgment. However, in the majority of cases, the bone age opinions diverged, and radiologists faced uncertainty in how to respond: "It [X-ray AI] would give me bone ages that would make me re-think what I said…I find that I'm often disagreeing with the model. Maybe it's just me and I don't know how to read bone ages" (Dr. D). Radiologists were troubled by the discrepancies, which led them to question their own judgments as well as the AI tool's, as in Dr. R remarking, "Sometimes I felt that the algorithm was a little inaccurate, either too old or too young…I couldn't put my finger on what it was that was off. Or maybe I was off, maybe the algorithm was more accurate, and I wasn't looking at it right."

Lacking the ability to understand or examine the tool's result left radiologists frustrated: "I have no idea, I really don't. I would be curious to know. I don't really know how it's working" (Dr. M). They were often questioning what the AI tool was considering and guessing at the image features the AI tool may be weighing: "I'd be curious to find out what parts of the image the algorithm actually uses… I felt it was probably looking at – I wasn't sure – but I felt like it was probably looking at more of the hand than I was … I don't know how much weight the AI gives to the different bones" (Dr. R). One afternoon, a spirited discussion broke out as Dr. D attempted to reason about the tool's underlying logic: "Is there a way to tell what the algorithm used on an individual case to come to its determination?...If it's looking at the wrist bones, we would maybe disagree with it." Although Dr. A agreed, she questioned Dr. D's assumption about how the tool was forming its judgment: "Yeah, but *is* it looking at the wrist bones?" Experiencing opacity of AI-in-use, Dr. D shrugged: "I don't know. I don't know how it works!" Dr. A sighed in frustration, agreeing: "It's a mystery."

In particular, they were baffled at how the AI tool was producing bone age measurements at a level of precision far greater than they were capable of producing. Pediatric radiologists report bone age results using the one-year increment standards, but the AI tool reported more granular results, using combinations of years and months (e.g., "6 years 4 months"), which Dr. R explained saying, "It [X-ray AI tool] doesn't always give you an exact number [of years]. It gives you a kind of interpolation between standards. We don't typically do that." They struggled to understand or interpret how the AI tool was able to discern these precise results that did not correspond with their accepted language or approach: "How is it coming up with this granular of a bone age? How does this make sense? How does it *know?*" (Dr. A).

Dealing with AI opacity: Not enacting AI interrogation practices and not incorporating AI results

As in the previous cases, pediatric radiologists encountered a sudden surge of uncertainty as they experienced opacity of AI-in-use. They struggled in the process of relating the AI tools' results to their own expert knowledge, as Dr. O remarked, "I don't really know how to gauge the results from that software; I'm not sure how it's working." Ultimately, in the cases we observed, pediatric radiologists did not enact AI interrogation practices and thus rarely incorporated AI results into their final judgments.

These radiologists faced a sudden increase in uncertainty when viewing the AI results despite the (previously) straightforward nature of the task. They were unable to integrate the tool's unfamiliar way of communicating bone age opinions with their own knowledge about pediatric bone development: "It [X-ray AI tool] gives me things like '11 years 8 months'. How does it get that?…If someone was going to ask me, 'How do you know it was 11 years 8 months?' I'd be like, 'I don't really know,'" (Dr. D). Moreover, they did not enact a rich range of analytical practices to help them interrogate the AI result and relate it to their own opinions. When they viewed a divergent AI bone age opinion, they resorted to re-reviewing the same images from the X-ray and textbook and rarely transformed their initial opinion as a result. This is illustrated in the following case. Dr. D's eyes flicked back and forth between the standard images and the patient's X-ray as she formed her initial assessment: "I'm looking at how wide is this area here [the areas separating the bones of the fingers]. Looking at the different shapes. This is bigger. This is the same. I think he's between 8 and 9. The machine says between 9 and 10. Closer to 10 actually!"

Reacting to the divergent AI opinion, Dr. D cocked her head to the side and exhaled in frustration: "Now I'm going to try to find *why* it said that." She continued reviewing the same image on her screen and the textbook again, which yielded no new insights that would change her original view: "I feel he's not that close to that [10 years]. I think the machine's overestimating. To me, it's 8 or 9."

DISCUSSION

Summary of findings

This study brings to light a process of how professional knowledge workers experienced and dealt with opacity of AI-in-use when forming critical judgments. In all three departments we studied, professionals' key practice is producing knowledge claims with the highest level of certainty possible. Professionals in two departments faced intense uncertainty (during lung cancer and breast cancer diagnosis) and worked hard to reduce using varied analytical practices. In the third department (when evaluating bone age), they experienced lower uncertainty and drew on fewer analytical practices. In all three departments, professionals first formed initial knowledge claims, then considered the AI knowledge claim, which frequently conflicted with their initial claim. In all three departments, professionals experienced opacity of AI-in-use, since they had no insight into the underlying reasoning of a given AI result, which in turn, heightened their experience of uncertainty.

Interestingly, the three departments had divergent patterns of the degree to which they transformed their own knowledge as a result of considering AI tool results. Only one department consistently integrated the AI results, despite the opacity of AI-in-use (when diagnosing lung cancer), whereas professionals in two other departments did not integrate the AI results (when diagnosing breast cancer and bone age). Upon closer analysis, we found it was critical that professionals were enacting "AI interrogation practices", or practices humans enact to relate their own knowledge claims to AI's knowledge claims. Enacting AI interrogation practices enabled professionals to reconcile the two knowledge claims (by overruling the AI claim, reflectively agreeing with it, or synthesizing the claims synergistically) and reduce their overall uncertainty. Professionals who did not enact such practices struggled to incorporate AI results due to the opacity and consequently formed their final judgments by

either blindly accepting or ignoring AI claims. We differentiate these paths of human-AI use as *engaged augmentation* and *un-engaged "augmentation."* Figure 1 summarizes this process and the two paths.

Insert Figure 1 about here

Theoretical and Practical Implications

Drawing on our conceptualization, we now outline the implications of our study for two key areas of focus for organizational scholars of AI: AI opacity and human-AI augmentation.

Opacity of AI-in-use and the importance of AI interrogation practices

Opacity associated with AI tools has become a fiercely debated topic in academic and societal conversations (Christin 2020, Diakopoulos 2020, von Krogh 2018, Pasquale 2015). Our study brings issues of opacity to the center stage in studying how professionals use AI tools to form critical judgments. Most of the existing literature on opacity conceptualizes opacity as a *property* of AI tools, especially of tools that use deep learning methods (Burrell 2016, Domingos 2015, Kellogg et al. 2019, Pearl and Mackenzie 2018). Our study shifts the analytical focus from what appears as an innate and fixed property of technology to the broader sociomaterial practice that *produces* opacity as a specific technology is used in a particular context. This enables us to focus on the process of how AI opacity *emerges* in practice and how, in some cases, professionals can deal with it.

A growing community focusing on issues of AI opacity proposes two approaches for dealing with it. The first focuses on limiting the use of AI tools for critical decisions if transparency is unattainable (e.g., Burrell 2016, Domingos 2015, Gillespie 2014, Teodorescu et al. 2021). The second approach is designing "explainable" or "interpretable" AI tools that provide greater transparency towards explaining AI outputs. While this work is critical (as we discuss below), our work uncovers a third approach. We illuminate a path where professionals deal with opacity of AI-in-use by enacting AI interrogation practices. These practices provided professionals a way of validating AI results, despite experiencing opacity, and resulted in an engaged mode of human-AI augmentation.

While many researchers are focused on developing "explainable AI" or "interpretable AI" (e.g., Barredo Arrieta et al. 2020, Bauer et al. 2021, Fernández-Loría et al. 2020, Guidotti et al. 2018, Hooker et al. 2019, Rudin 2019, Samek et al. 2019, Teodorescu et al. 2021), some leading scholars (Cukier et al. 2021, Simonite 2018) and AI designers believe there is no need for explanations. They argue that an AI tool's evidence-based performance results should motivate experts to rely on the tool's results with confidence. This assumption was expressed by a leader of AI research at Urbanside: "People talk about explainability in AI a lot. My personal opinion is I don't think you need to do *any* explaining. As long as you show users that the tool performs well. When it performs well, I think people are really okay working under that uncertainty." Our study shows how that point of view is disconnected from the reality of how professionals are wrestling with opacity of using AI in practice. Based on our study's findings, explainable or interpretable AI may enable but not guarantee that professionals are able to integrate AI knowledge (i.e., engaged augmentation). Our study showed that despite AI tools' high performance documented in published literature, some professionals chose to invest their valuable time into AI interrogation practices rather than simply relying on unexplained AI claims at face value.

If new generation AI tools provide explanations or become more interpretable, this should impact experts' *ability* to engage with AI, but not necessarily their *motivation* or *willingness* to do so. Such willingness is influenced by many factors such as professional norms, organizational and financial incentives, and societal expectations. Making sense of AI explanations requires an investment of time and resources. Not only is this challenging given intense organizational constraints (e.g., time, knowledge), but such investment does not align with widely held expectations that AI will make work faster and more efficient. In medical practice, professionals solicit opinions from their colleagues, investing in collaboration only when they experience particularly high doubt or uncertainty (on a regular, but infrequent basis). In contrast, in our study (as in many leading US hospitals) the AI tool provides opinions on every case, regardless of the professional's degree of uncertainty. Thus, professionals were spending additional time coping with the heightened uncertainty, even for simple and routine cases (where promises of AI efficiency are strongest). We hope future research will further unpack the relationship

between AI and time as the push for accelerating the pace of work is increasing (Lifshitz-Assaf et al. 2021), yet implications on the nature and quality of work are underexplored.

Our study also contributes to the debates and conversations on opacity by uncovering an important relationship between opacity of AI-in-use and epistemic uncertainty (Griffin and Grote 2020, Packard and Clark 2020, Rindova and Courtney 2020). In many knowledge fields, experts are keenly focused on producing high-quality judgments and they are willing to invest resources to obtain additional evidence and reduce their epistemic uncertainty – or ignorance of unknown but knowable information (Packard and Clark 2020). Contrary to prior literature, when professionals in our study obtained additional "evidence" from AI tools which often diverged from their prior judgment, their epistemic uncertainty increased due to their experience of opacity. In our study, professionals would regularly integrate conflicting knowledge provided by their colleagues by probing one another and building on their common ground and participation in a shared field (Carlile 2004, Levina 2005, Maguire et al. 2004). However, when professionals' opinions diverged from AI tools' opinions, no common ground or shared field exists or can be created (as tools are designed today). Enacting AI interrogation practices was the only way some professionals were able to overcome the opacity of AI-in-use and reduce the uncertainty needed to integrate the AI knowledge into their own.

Future research is warranted to explain why some professionals enact AI interrogation practices while others do not. Our study suggests three main potential factors: the AI tool's ability to reduce professionals' uncertainty, the presence of time pressure (and other resource constraints) on professionals' work, and the richness of professionals' complementary technologies-in-use. Motivation to invest in AI interrogation practices may be lower if professionals view the AI expertise as similar to (or worse than) their own. In such cases, there is only increased pressure of investing additional time without the benefit of reducing uncertainty (as in the breast and pediatric departments in our study). Moreover, the time required to enact AI interrogation practices may further deter professionals from investing in them (as in the breast department, where time pressure was extremely high). It is also possible that professionals may still develop AI interrogation practices as they continue using the AI tool over a longer period of time (as

in the pediatric department); on the other hand, it may be difficult to develop such practices when the complementary technologies are limited and lack richness (e.g., when analyzing X-ray images). Future research should investigate other motivators or deterrents that were not apparent in our context, such as the impact of regulation or perceived legal and institutional risks. It could be, for example, that regulatory or authority bodies that require professionals to articulate why they overrule an AI result may motivate investment in AI interrogation practices.

Importantly, we do not wish to suggest that AI interrogation practices are a substitute for explainability or interpretability. On the contrary, we urge continued dedicated attention and resources towards designing AI tools that enable professionals to more readily integrate AI knowledge claims in practice. For instance, when investigating the X-ray AI tool for determining bone age, we, as academic researchers learned from reading archival published materials that it is possible to produce salience maps showing what areas on the X-ray were most relevant for producing a given AI result. While this technology is available for the algorithm underlying the X-ray tool, it was not implemented at Urbanside. By highlighting the critical importance of AI interrogation practices, we hope that managers, practitioners, and researchers will focus on designing and adopting more transparent and interpretable AI tools. This should help professionals to more easily develop AI interrogation practices that build on these explanations. In this study, all AI tools had a similar degree of opacity of AI-in-use. Future studies can explore a variety of interrogation practices that may emerge in response to different degrees of opacity. Moreover, AI researchers can proactively design new features that ease engagement.

We also wish to highlight our focus on professionals' judgments for *critical* decisions, those with particularly high consequences or costs of errors (medical diagnoses in our study). Our findings are relevant to contexts where experts make critical decisions that require knowledge integration and transformation, such as judges rendering verdicts and sentencing, human resource managers evaluating employees, or military experts carrying out targeted attacks. Our study speaks to such contexts where engaged augmentation is necessary versus those where experts may defer to AI results even when opaque. We do not suggest our findings apply for decisions that do not require knowledge transformation or when

the cost of errors is substantially low, such as using AI for supply chain logistics, marketing and advertising, grammatical editing, or call center prioritization.

Future research is needed to explore potential differences in how professionals in other contexts experience opacity and enact AI interrogation practices. This is a study of a specific profession (physicians) within a highly-resourced US organization (a teaching hospital), and we believe other experts in different legal and professional environments are important to investigate. The organization we studied has world-leading experts with high standards of quality and strong professional accountability. In the past few years, there has been a gold rush to purchasing AI tools, especially in hospitals with fewer resources and lower standards of care (Gkeredakis et al. 2021, Moran 2018, Roberts et al. 2021). Based on this study, we suggest that such a gold rush may give rise to un-engaged "augmentation" which is highly risky from a learning and knowledge perspective for experts, AI companies, and consumers. In addition, our study is based in the US legal system whereby hospitals must adhere to strict regulation and oversight, which is not the case in many countries currently adopting AI tools. Future research is warranted on the role of regulation on the adoption and engagement of AI tools for critical decisions. When such regulation is missing, and fewer checks and validations are in place, engaged augmentation may be even less likely, and yet even more important.

Challenging the taken-for-granted concept of augmentation

Professional work is currently being disrupted by AI technologies, as modern AI increasingly pertains to processes of producing and evaluating knowledge claims (Anthony 2018, Faraj et al. 2018, von Krogh 2018, Pachidi et al. 2020). Debates are emerging around the degree of automation or augmentation that may result as AI tools are adopted into professional work settings (e.g., Autor 2015, Seamans and Furman 2019, Zhang et al. 2021). Our study speaks to this important debate by problematizing the taken-for-granted concept of augmentation and its implication for the future of work and human expertise.

Within the current literature, augmentation generally refers to human in the loop scenarios where experts and AI tools "collaborate" so as to "*multiply" and* "*combine their complementary strengths*" (Raisch and Krakowski 2021, p. 6). The results of our study challenge the taken-for-granted equivalency

of augmentation with collaboration. Instead, we suggest differentiating *engaged augmentation* from *un-engaged "augmentation."* In engaged augmentation, experts integrated AI knowledge claims with their own, which requires both building an understanding of the AI claim and the ability and willingness to transform one's own knowledge based on the AI claim (this took place in lung cancer diagnosis in our study). By enacting AI interrogation practices, professionals were able to understand the AI result, albeit the opacity they experienced, and demonstrated their willingness to change their initial judgment through reflectively agreeing with the AI claim, overruling it, or synthesizing the two claims. From a learning and knowledge perspective, engaged augmentation scenarios could be productive and beneficial, including cases of reflectively overruling the AI results. Future research is needed to investigate the learning that professionals (and AI tools) experience when involved in engaged augmentation over extended periods of time. For example, if engaged professionals routinely change their judgment by endorsing AI results, they may be reproducing AI's shortcomings over time (e.g., AI errors or biases in judgment).

In contrast, un-engaged "augmentation" involved professionals not relating the AI knowledge claims to their own claims (this took place during breast cancer and bone age diagnosis in our study). These professionals appeared to be using the AI tool as they were going through the act of opening the AI results. However, they were not integrating knowledge claims and mostly blindly accepting or blinding ignoring the AI results. This path does not offer strong opportunities or benefits from a learning and knowledge perspective. We argue that human in the loop scenarios of un-engaged "augmentation" are essentially cases of automation: without the ability to relate AI knowledge claims to experts' (through AI interrogation practices or explainable or interpretable AI tools that enable interrogation), what looks like augmentation on paper is much closer to automation.

Moreover, there is an assumption that augmentation will help organizations achieve positive outcomes, usually through improving humans' knowledge insights, efficiency, or both (Brynjolfsson and Mitchell 2017, Daugherty and Wilson 2018, Davenport and Kirby 2016, Raisch and Krakowski 2021). Researchers have already raised issues of AI accuracy claims and "superior to human" knowledge performance (Lebovitz et al. 2021). Using AI tools *may* have reduced some human error but could have

introduced other errors into the human's judgment (e.g., due to biases or poor training data). To truly understand the impact on accuracy, we must be able to compare AI outputs and experts' judgments. However, in many professional contexts, such evaluations are limited since the knowledge is highly uncertain and many "ground truth" measures are based on knowledge claims that lack strong external validation (Lebovitz et al. 2021).

Our study adds to these concerns by calling into question the assumption of increased efficiency. In all three cases we studied, experts using AI tools spent additional time even on "simple" cases as they experienced opacity and additional uncertainty. Within engaged augmentation, experts invested additional time to reconcile the AI knowledge claims by enacting AI interrogation practices. This additional time may be justified by improvements to care quality, but it is unclear whether healthcare systems are willing to commit that additional time. AI vendors tend to promote their tools using promises of efficiency. If managers implement these tools based on such claims, they may pressure experts to reduce time spent per judgment, which is likely to encourage un-engaged "augmentation" and potentially lead to a decline in quality (e.g., patient health outcomes).

Another perspective on our study that warrants future research is the impact of AI on an overall professional field and its knowledge work over time. Leading medical professionals have been claiming that AI tools will eliminate the need for professional radiologists, explicitly citing the rise of diagnostic AI tools as a case for automation. On the other hand, leading radiologists are arguing that AI can enhance their professional roles and abilities. In our study, we did not find significant differences in attitudes towards AI across departments, which all took positive approaches towards adopting new technologies (including AI). Future research may explore the impact of AI on the broader professional field of radiology and other professions experiencing massive disruption due to AI tools (e.g., HR, criminal justice). It could be that engaged augmentation or un-engaged "augmentation" are reactive responses of professionals dealing with the potential disruption posed by AI and automation. It will be important to investigate how professionals respond to a new technological force that is challenging the professional jurisdiction and knowledge boundaries of an existing profession, for instance, how professionals enact

professional identity work (Lifshitz-Assaf 2018, Tripsas 2009) or knowledge boundary work (Barrett et al. 2012, Bechky 2003b, Levina and Vaast 2005) or the strategies and responses that impact the profession field (Bechky 2019, Howard-Grenville et al. 2017, Nelson and Irwin 2014).

To conclude, we do not wish to convey that dealing with the opacity related to AI tools, even by using AI interrogation practices, should be viewed as the desired or optimal path forward. From knowledge and learning perspectives, the opacity of AI tools can be seen as inhibiting knowledge workers' full feedback and reflective cycles (Gherardi 2000, Schön 1983). When professionals cannot analyze the reasoning behind AI decisions, they miss out on the learning process (Beer 2017), lacking opportunities to reflect on, deepen, or update their expertise (Beane 2019). Ultimately, AI technologies are designed to create new sorts of expertise to enable professionals, organizations, and even society to better address hard problems such as medical diagnosis. However, the opacity experienced when professionals are using AI tools is an increasingly critical problem in its own right. We urge further researchers and policy-makers to tackle this problem, across domains and disciplines, to ensure that the path of new technological development meets the needs of humanity and society.

REFERENCES

Albu OB, Flyverbom M (2019) Organizational Transparency: Conceptualizations, Conditions, and Consequences. *Business & Society* 58(2):268–297.

Ananny M, Crawford K (2016) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20(3):973–989.

Anthony C (2018) To Question or Accept? How Status Differences Influence Responses to New Epistemic Technologies in Knowledge Work. *AMR* 43(4):661–679.

Anthony C (2021) When Knowledge Work and Analytical Technologies Collide: The Practices and Consequences of Black Boxing Algorithmic Technologies. *Administrative Science Quarterly*:00018392211016755.

Autor DH (2015) Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *The Journal of Economic Perspectives* 29(3):3–30.

Bailey D, Leonardi P, Barley S (2012) The Lure of the Virtual. *Organization Science* 23(5):1485–1504.

Barad K (2003) Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society* 28(3):801–831.

Barley S (1986) Technology as an occasion for structuring: Technically induced change in the temporal organization of radiological work. *Administrative Science Quarterly* 3(1):78–108.

Barley S (1990) The alignment of technology and structure through roles and networks. *Administrative Science Quarterly* 35(1):61–103.

Barley SR, Bechky BA, Milliken FJ (2017) The Changing Nature of Work: Careers, Identities, and Work Lives in the 21st Century. *Academy of Management Discoveries* 3(2):111–115.

Barocas S, Selbst AD, Raghavan M (2020) The hidden assumptions behind counterfactual explanations and principal reasons. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. (Association for Computing Machinery, New York, NY, USA), 80–89.
Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, et al. (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58:82–115.
Barrett M, Oborn E, Orlikowski W (2016) Creating Value in Online Communities: The Sociomaterial Configuring of Strategy, Platform, and Stakeholder Engagement. *Information Systems Research* 27(4):704–723.
Barrett M, Oborn E, Orlikowski WJ, Yates J (2012) Reconfiguring Boundary Relations: Robotic Innovations in Pharmacy Work. *Organization Science* 23(5):1448–1466.
Bauer K, Hinz O, van der Aalst W, Weinhardt C (2021) Expl(AI)n It to Me – Explainable AI and Information Systems Research. *Bus Inf Syst Eng* 63(2):79–82.
Beane M (2019) Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail. *Administrative Science Quarterly* 64(1):87–123.
Beane M, Orlikowski WJ (2015) What Difference Does a Robot Make? The Material Enactment of Distributed Coordination. *Organization Science* 26(6):1553–1573.
Bechky B (2003a) Sharing Meaning Across Occupational Communities: The Transformation of Understanding on a Production Floor. *Organization Science* 14(3):312–330.
Bechky B (2003b) Object lessons: Workplace artifacts as representations of occupational jurisdiction. *American Journal of Sociology* 109(3):720–752.
Bechky BA (2019) Evaluative Spillovers from Technological Change: The Effects of "DNA Envy" on Occupational Practices in Forensic Science. *Administrative Science Quarterly*.
Beer D (2017) The social power of algorithms. *Information, Communication & Society* 20(1):1–13.
Benbya H, Pachidi S, Jarvenpaa S (2021) Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems* 22(2).
Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K (2020) Fairlearn: A toolkit for assessing and improving fairness in AI.
Boyaci T, Canyakmaz C, deVericourt F (2020) *Human and Machine: The Impact of Machine Input on Decision-Making Under Cognitive Limitations* (Social Science Research Network, Rochester, NY).
Brynjolfsson E, McAfee A (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (W. W. Norton & Company, New York, NY).
Brynjolfsson E, Mitchell T (2017) What can machine learning do? Workforce implications. *Science* 358(6370):1530–1534.
Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1).
Caplan R, Donovan J, Hanson L, Matthews J (2018) *Algorithmic accountability: A primer* (Data & Society).
Carlile PR (2004) Transferring, Translating, and Transforming: An Integrative Framework for Managing Knowledge Across Boundaries. *Organization Science* 15(5):555–568.
Charmaz K (2014) *Constructing Grounded Theory* (Sage, Thousand Oaks, CA).
Christin A (2020) The ethnographer and the algorithm: beyond the black box. *Theor Soc* 49(5):897–918.
Crawford K, Dobbe R, Dyer T, Fried G, Green B, Kaziunas E, Kak A, et al. (2019) *AI Now 2019 Report* (AI Now Institute, New York, NY).
Cremer DD, Kasparov G (2021) AI Should Augment Human Intelligence, Not Replace It. *Harvard Business Review* (March 18) https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it.
Cukier K, Mayer-Schonberger V, De Vericourt F (2021) *FRAMERS: HUMAN ADVANTAGE IN AN AGE OF TECHNOLOGY AND TURMOIL* (Dutton).

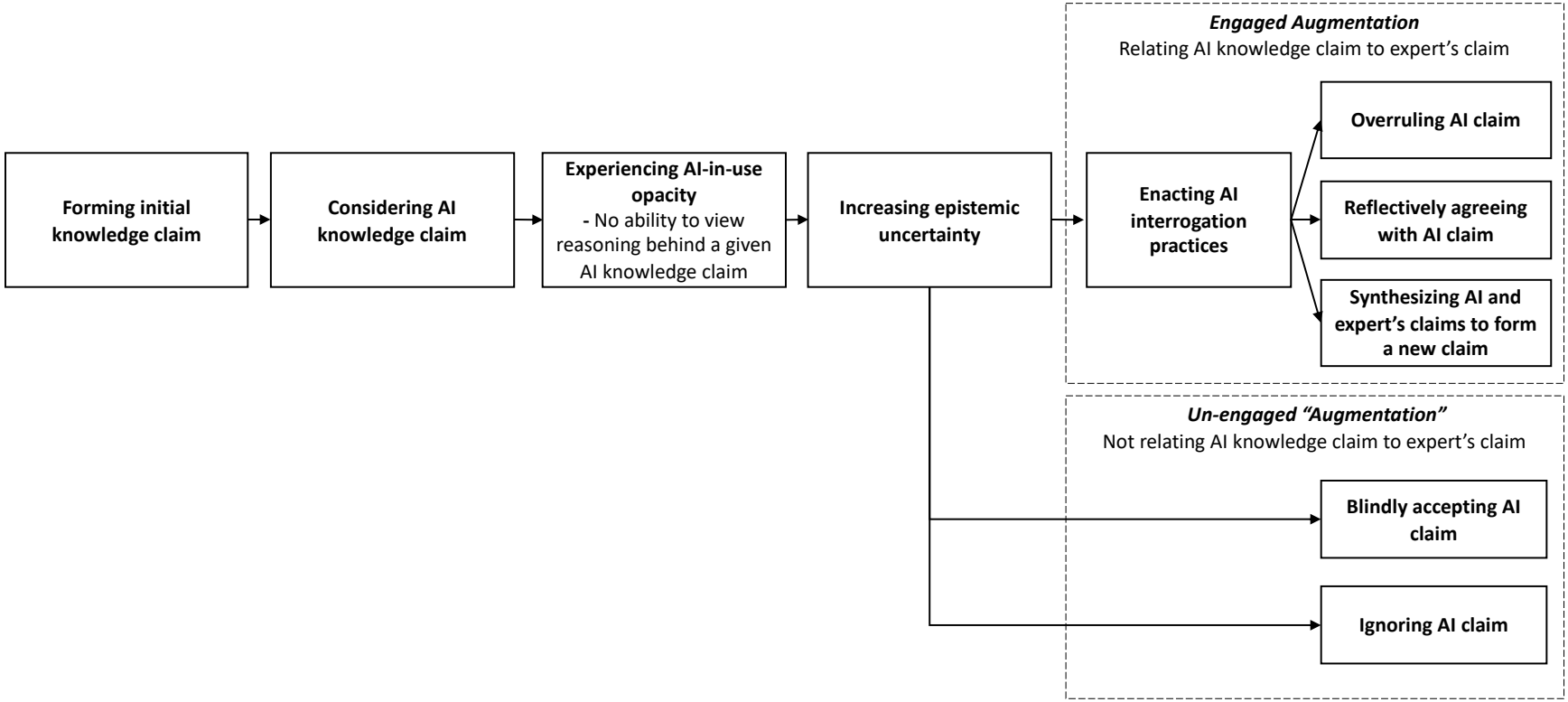
Daugherty PR, Wilson HJ (2018) *Human + Machine: Reimagining Work in the Age of AI* (Harvard Business Press).
Davenport TH, Kirby J (2016) *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines* (HarperBusiness, New York, NY).
Diakopoulos N (2020) Transparency. Dubber M, Pasquale F, Das S, eds. *The Oxford Handbook of Ethics of AI*. (Oxford University Press), 197–214.
Dodgson M, Gann DM, Salter A (2007) "In Case of Fire, Please Use the Elevator": Simulation Technology and Organization in Fire Engineering. *Organization Science* 18(5):849–864.
Domingos P (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* 1 edition. (Basic Books).
Dourish P (2016) Algorithms and their others: Algorithmic culture in context. *Big Data & Society* 3(2).
Erickson I, Robert L, Nickerson J (2018) Workshop: Work in the Age of Intelligent Machines. *GROUP '18 Proceedings of the 2018 ACM Conference on Supporting Groupwork Pages 359-361*.
Faraj S, Pachidi S, Sayegh K (2018) Working and organizing in the age of the learning algorithm. *Information and Organization* 28(1):62–70.
Fernández-Loría C, Provost F, Han X (2020) Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach. *arXiv:2001.07417 [cs, stat]*.
Gao R, Saar-Tsechansky M, De-Arteaga M, Han L, Lee MK, Lease M (2021) Human-AI Collaboration with Bandit Feedback. *arXiv:2105.10614 [cs]*.
Gherardi S (2000) Practice-Based Theorizing on Learning and Knowing in Organizations. *Organization* 7(2):211–223.
Gillespie T (2014) The Relevance of Algorithms. Gillespie T, Boczkowski PJ, Foot KA, eds. *Media Technologies: Essays on Communication, Materiality, and Society*. (MIT Press), 167–194.
Gkeredakis M, Lifshitz-Assaf H, Barrett M (2021) Crisis as opportunity, disruption and exposure: Exploring emergent responses to crisis through digital technology. *Information and Organization*:100344.
Glaser B, Strauss A (1967) *Discovering Grounded Theory* (Aldine Publishing Company, Chicago, IL).
Glikson E, Woolley AW (2020) Human Trust in Artificial Intelligence: Review of Empirical Research. *ANNALS* 14(2):627–660.
Grady D (2019) A.I. Took a Test to Detect Lung Cancer. It Got an A. *The New York Times* (May 20) https://www.nytimes.com/2019/05/20/health/cancer-artificial-intelligence-ct-scans.html.
Griffin M, Grote G (2020) When is more uncertainty better? A model of uncertainty regulation and effectiveness. *Academy of Management Review* 45(4):745–765.
Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51(5):93:1-93:42.
Hansen HK, Flyverbom M (2015) The politics of transparency and the calibration of knowledge in the digital age. *Organization* 22(6):872–889.
Haraway D (1988) Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3):575–599.
Hardy C, Lawrence TB, Grant D (2005) Discourse and Collaboration: The Role of Conversations and Collective Identity. *AMR* 30(1):58–77.
Hooker S, Erhan D, Kindermans PJ, Kim B (2019) A benchmark for interpretability methods in deep neural networks. 9737–9748.
Howard-Grenville J, Nelson AJ, Earle A, Haack J, Young D (2017) "If Chemists Don't Do It, Who Is Going To?" Peer-driven Occupational Change and the Emergence of Green Chemistry. *Administrative Science Quarterly* 62(3):524–560.
Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (2020) Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. (Association for Computing Machinery, New York, NY, USA), 1–14.

Kellogg K, Valentine M, Christin A (2019) Algorithms at work: The new contested terrain of control. *Academy of Management Annals*.
Khadpe P, Krishna R, Fei-Fei L, Hancock JT, Bernstein MS (2020) Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4(CSCW2):163:1-163:26.
Kogut B, Zander U (1992) Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. *Organization Science* 3(3):383–397.
von Krogh G (2018) Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *AMD* 4(4):404–409.
Latour B (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory* (Oxford University Press., Oxford, U.K.).
Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly* 45(3):1501–1525.
Leonardi P (2011) When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly* 35(1):147–167.
Leonardi P, Barley S (2010) What's under construction here? Social action, materiality, and power in constructivist stdies of technology and organizing. *Academy of Management Annals* 4(1):1–51.
Leonardi PM, Bailey DE (2008) Transformational Technologies and the Creation of New Work Practices: Making Implicit Knowledge Explicit in Task-Based Offshoring. *MIS Quarterly* 32(2):411–436.
Leonardi PM, Treem JW (2020) Behavioral Visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies* 41(12):1601–1625.
Levina N (2005) Collaborating on Multiparty Information Systems Development Projects: A Collective Reflection-in-Action View. *Information Systems Research* 16(2):109–130.
Levina N, Vaast E (2005) The Emergence of Boundary Spanning Competence in Practice: Implications for Implementation and Use of Information Systems. *MIS Quarterly* 29(2):335–363.
Lifshitz-Assaf H (2018) Dismantling Knowledge Boundaries at NASA: The Critical Role of Professional Identity in Open Innovation. *Administrative Science Quarterly* 63(4):746–782.
Lifshitz-Assaf H, Lebovitz S, Zalmanson L (2021) Minimal and Adaptive Coordination: How Hackathons' Projects Accelerate Innovation without Killing it. *AMJ* 64(3):684–715.
Maguire S, Hardy C, Lawrence TB (2004) Institutional Entrepreneurship in Emerging Fields: HIV/AIDS Treatment Advocacy in Canada. *AMJ* 47(5):657–679.
Mazmanian M, Cohn M, Dourish P (2014) Dynamic Reconfiguration in Planetary Exploration: A Sociomaterial Ethnography. *MIS Quarterly* 38(3):831–848.
Mazmanian M, Orlikowski W, Yates J (2013) The Autonomy Paradox: The Implications of Mobile Email Devices for Knowledge Professionals. *Organization Science* 24(5):1337–1357.
Mol A (2003) *The Body Multiple* (Duke Universoty Press).
Moran G (2018) This artificial intelligence won't take your job, it will help you do it better. *Fast Company* (October 24) https://www.fastcompany.com/90253977/this-artificial-intelligence-wont-take-your-job-it-will-help-you-do-it-better.
Mukherjee S (2017) A.I. Versus M.D. (March 27) https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.
Nelson AJ, Irwin J (2014) "Defining What We Do—All Over Again": Occupational Identity, Technological Change, and the Librarian/Internet-Search Relationship. *AMJ* 57(3):892–928.
Nunn J (2018) How AI Is Transforming HR Departments. *Forbes* https://www.forbes.com/sites/forbestechcouncil/2018/05/09/how-ai-is-transforming-hr-departments/.
Orlikowski W (1992) The duality of technology: Rethinking the concept of technology in organizations. *Organization Science* 3(3):398–427.
Orlikowski W (2000) Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science* 11(4):404–428.
Orlikowski W (2007) Sociomaterial Practices: Exploring Technology at Work. *Organization Studies* 28(9):1435–1448.

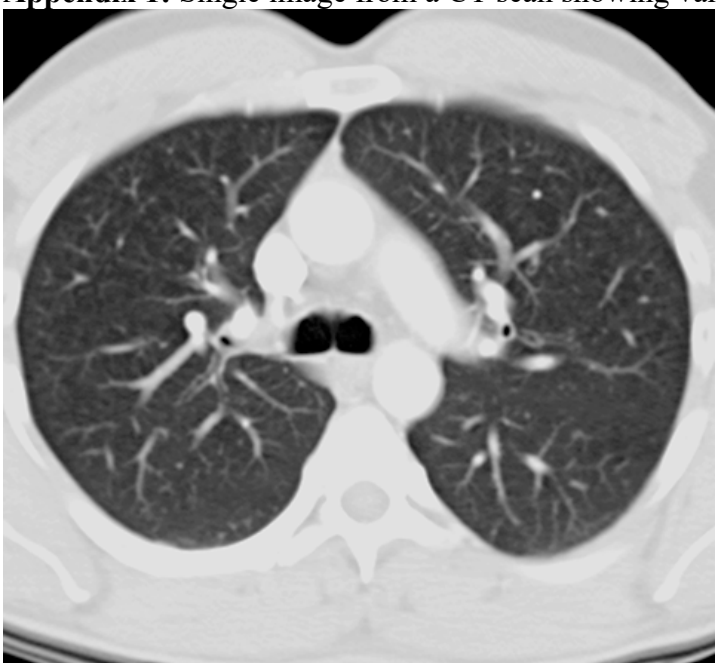
Orlikowski W, Scott S (2008) Sociomateriality: Challenging the Separation of Technology, Work and Organization. *Academy of Management Annals* 2(1):433–474.
Pachidi S, Berends H, Faraj S, Huysman M (2020) Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing. *Organization Science*.
Packard MD, Clark BB (2020) On the Mitigability of Uncertainty and the Choice between Predictive and Nonpredictive Strategy. *AMR* 45(4):766–786.
Pasquale F (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information* Reprint edition. (Harvard University Press, Cambridge, Massachusetts London, England).
Pearl J, Mackenzie D (2018) *The Book of Why: The New Science of Cause and Effect* 1st edition. (Basic Books, New York).
Pinch T, Bijker W (1987) The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. Hughes TP, Bijker W, Pinch T, eds. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. (MIT Press), 17–50.
Polanyi M (1958) *Personal Knowledge: Towards a Post-Critical Philosophy* (University of Chicago Press, Chicago).
Polanyi M (1966) *The Tacit Dimension* (University of Chicago Press, Chicago ; London).
Puranam P (2021) Human–AI collaborative decision-making as an organization design problem. *J Org Design*.
Rai A, Constantinides P, Sarker S (2019) Editor's Comments: Next-Generation Digital Platforms: Toward Human–AI Hybrids. *Management Information Systems Quarterly* 43(1):iii–ix.
Raisch S, Krakowski S (2021) Artificial Intelligence and Management: The Automation–Augmentation Paradox. *AMR* 46(1):192–210.
Razorthink Inc. (2019) 4 Major Challenges facing Fraud Detection; Ways to Resolve Them using Machine Learning. *Medium* (April 25) https://medium.com/razorthink-ai/4-major-challenges-facing-fraud-detection-ways-to-resolve-them-using-machine-learning-cf6ed1b176dd.
Recht M, Bryan RN (2017) Artificial Intelligence: Threat or Boon to Radiologists? *Journal of the American College of Radiology* 14(11):1476–1480.
Rindova V, Courtney H (2020) To Shape or Adapt: Knowledge Problems, Epistemologies, and Strategic Postures under Knightian Uncertainty. *AMR* 45(4):787–807.
Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, Aviles-Rivero AI, et al. (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3(3):199–217.
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1(5):206–215.
Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR (2019) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer Nature).
Schön DA (1983) *The Reflective Practitioner : How Professionals Think in Action* (Basic Books, New York, NY).
Scott S, Orlikowski W (2014) Entanglements in practice: Performing anonymity through social media. *MIS Quarterly* 38(3):873–893.
Scott SV, Orlikowski WJ (2012) Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, Organizations and Society* 37(1):26–40.
Seamans R, Furman J (2019) AI and the Economy. *Innovation Policy and the Economy* 19(1):161–191.
Simonite T (2018) Google's AI Guru Wants Computers to Think More Like Brains. *Wired Magazine* (December 12) https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/.
Stohl C, Stohl M, Leonardi PM (2016) Digital Age | Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age. *International Journal of Communication* 10(0):15.
Suchman L (2007) *Human-Machine Reconfigurations: Plans and Situated Actions* (Cambridge University Press).

Teodorescu M, Morse L, Awwad Y, Kane G (2021) Failures of Fairness in Automation Require a Deeper Understanding of Human–ML Augmentation. *MIS Quarterly*.
Tripsas M (2009) Technology, Identity, and Inertia Through the Lens of "The Digital Photography Company." *Organization Science* 20(2):441–460.
Turco CJ (2016) *The Conversational Firm: Rethinking Bureaucracy in the Age of Social Media* (Columbia University Press).
Van Den Broek E, Sergeeva A, Huysman M (2021) When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS Quarterly* 45(3):1557–1580.
Van Maanen J (1988) *Tales of the Field: On Writing Ethnography, Second Edition* (University of Chicago Press, Chicago, IL).
Waardenburg L, Sergeeva A, Huysman M (2018) Hotspots and Blind Spots. Schultze U, Aanestad M, Mähring M, Østerlund C, Riemer K, eds. *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology*. IFIP Advances in Information and Communication Technology. (Springer International Publishing, Cham), 96–109.
Wagner E, Newell S, Piccoli G (2010) Understanding Project Survival in an ES Environment: A Sociomaterial Practice Perspective. *Journal of the Association for Information Systems* 11(5):276–297.
Wagner EL, Moll J, Newell S (2011) Accounting logics, reconfiguration of ERP systems and the emergence of new accounting practices: A sociomaterial perspective. *Management Accounting Research* 22(3):181–197.
Watkins EA (2020) Took a Pic and Got Declined, Vexed and Perplexed: Facial Recognition in Algorithmic Management. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. CSCW '20 Companion. (Association for Computing Machinery, New York, NY, USA), 177–182.
Wilson HJ, Daugherty PR (2018) Collaborative Intelligence: Humans and AI Are Joining Forces. *Harvard Business Review* (July 1) https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces.
Xu F, Uszkoreit H, Du Y, Fan W, Zhao D, Zhu J (2019) Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. Tang J, Kan MY, Zhao D, Li S, Zan H, eds. *Natural Language Processing and Chinese Computing*. Lecture Notes in Computer Science. (Springer International Publishing, Cham), 563–574.
Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, et al. (2021) *The AI Index 2021 Annual Report* (AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA).
Zuboff S (2015) Big other: surveillance capitalism and the prospects of an information civilization. *J Inf Technol* 30(1):75–89.

Figure 1: Experts using AI tools for critical judgments

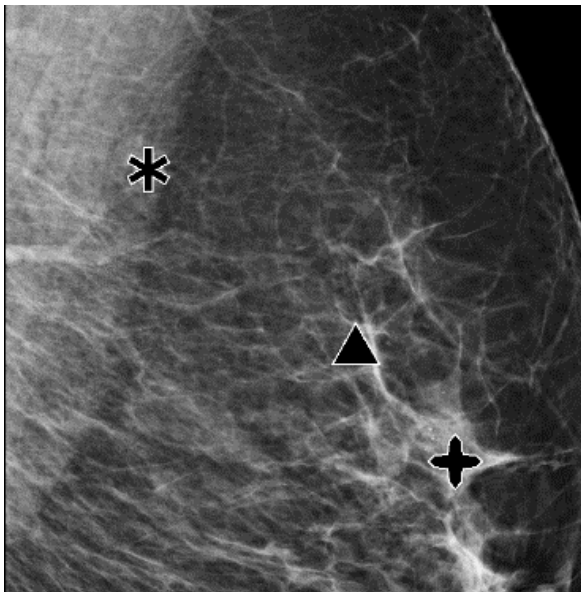


Appendix 1: Single image from a CT scan showing various lung structures



Appendix 2: Typical display of digital mammogram image on radiologists' workstation

Appendix 3: Mammo AI tool outputs



Appendix 4: Typical display of digital hand X-ray used for bone age diagnosis