Association for Information Systems

AIS Electronic Library (AISeL)

ICIS 2019 Proceedings

The Future of Work

Diagnostic Doubt and Artificial Intelligence: An Inductive Field Study of Radiology Work

Sarah Lebovitz
New York University, sarah.lebovitz@stern.nyu.edu

Diagnostic Doubt and Artificial Intelligence: An Inductive Field Study of Radiology Work

Completed Research Paper

Sarah Lebovitz
New York University
44 W 4th Street New York, NY
Sarah.lebovitz@stern.nyu.edu

Abstract

Technological developments in emerging AI technologies are assumed to further routinize and improve the efficiency of decision making tasks, even in professional contexts such as medical diagnosis, human resource management, and criminal justice. We have little research on how AI technologies are actually used and adopted in practice. Prior research on technology in organizations documents a gap between the expectations for new technology and its actual use in practice. We conducted a comparative field study of three sections in a Department of Radiology in a major US hospital, whereby new and existing AI tools were being used and experimented with. In contrast to expectations about AI tools, our study reveals how such tools can lead routine professional decision making tasks to become nonroutine, as they increased ambiguity and decision makers had to work to reduce it. This is particularly challenging since the costs of dealing with ambiguity – increased time to diagnose – were often weighed against the benefits of such ambiguity (potentially more accurate diagnoses). This study contributes to literatures related to technology, work, and organizations, as well as the role of ambiguity in professionals' knowledge work.

Keywords: Artificial intelligence, professionalism, new technology and work, innovation

Introduction

Technological developments in Artificial Intelligence (AI) promise continuous improvement in decision-making, problem-solving, and reasoning capabilities that edge closer and closer to human capabilities (Davenport and Kirby 2016). Augmenting and automating these tasks with AI may fundamentally shift the nature and prevalence of knowledge work in future organizations and reimagine how knowledge is produced and evaluated by new configurations of humans and machines (Erickson et al. 2018; Faraj et al. 2018). In this context, AI generally refers to a wide range of algorithms and computer-aided systems that detect patterns among data sets of past behaviors and link these patterns to potential future outcomes (von Krogh 2018). AI adoption falls within the larger phenomenon of digital transformation, which has become a central theme in how we theorize myriad dimensions of work and organizational life (Bailey et al. 2012; Kellogg et al. 2006; Orlikowski 2010; Valentine et al. 2017; Zammuto et al. 2007).

While AI is within the realm of digital transformations, its particular characteristics have elicited an intense response from practitioners and scholars alike. Yet amidst this exploding discourse, we lack a deep understanding of how AI impacts individuals, groups, and organizational outcomes in practice. Scholars of technology in organizations are urged to study this "new and poorly understood phenomenon" (von Krogh 2018, p. 408), and, specifically, the ways in which AI augments diverse dimensions of organizational life. Even organizational creativity scholars (Amabile 2019) are called upon to examine new configurations of human-machine intelligence and their effects on the production of new and creative ideas. This study follows recent calls for phenomenon-based examination of AI in organizational life and investigates four implementations of AI in a radiology department.

The field of diagnostic radiology is expected to undergo dramatic transformation in the age of AI. Leading experts in artificial technology, such as Professor Geoffrey Hinton, stated that people should stop training radiologists as AI made technological leaps in image processing capabilities. Radiology is at the cutting-edge of adopting emerging technologies, including AI, which makes this setting particularly relevant for learning about the unfolding nature of AI use in professional knowledge work for high judgment decisions. This paper is based on a field study of three sections within a department of radiology using different AI tools at a major tertiary hospital in the United States. This study reveals how, even for routine tasks, AI induces increased diagnostic ambiguity and doubt as radiologists struggle to produce accurate and time-sensitive medical diagnoses using such tools. This is critical as radiologists must weigh the costs and benefits of using AI in their decision making process. This study contributes to the nascent understanding of the adoption and use of emerging intelligent technologies in professional knowledge work and to our understanding of ambiguity in such work.

Theoretical Perspectives

Artificial Intelligence in Organizations

The term “artificial intelligence” and its associated capabilities have transformed over the past decades. Over four decades ago, Herbert Simon argued that “every manager needs to be able to analyze problems systematically (and with the aid of the modern arsenal of analytical tools provided by management science and AI)” (1987, p. 63). This iteration of “artificial intelligence” refers to early computers that performed computationally heavy tasks like census tabulation or weather pattern analysis. Today, the term “artificial intelligence” is associated with the ability of technology to make predictions about the future (Chen et al. 2012). Advances in AI stem from the exponential availability of data capturing diverse dimensions of organizational and private life as well as technological developments in computer science and engineering to support efficient and low-cost data processing (Chen et al. 2012; von Krogh 2018; Lycett 2013).

Following these continuous technological advances, organizational scholars are beginning to take notice of the impacts of AI on organizational life. Speculating about the diverse ways AI may transform organizations, Faraj and colleagues (2018) call for studies where AI is implemented to accomplish work previously involving human input and reasoning. Today, radiology work is highly subject to human interpretation and reasoning. Automating such work raises questions about the ability of AI to mimic or surpass the quality of decisions made by humans. Von Krogh (2018) offers useful direction for this area of research. He suggests that studies of AI should consider task inputs, task process, and task output to understand the purpose and functioning of AI in relation to humans in organizational settings. This study follows his call to examine the meanings and uses of AI with respect to specific features of the decision-making tasks.

Organizational scholars have begun looking at algorithms and the resulting the reconfiguration of practices in contexts such as sales (Pachidi et al. 2019), travel and hospitality evaluation (Orlikowski and Scott 2014), ride-sharing (Cameron Working Paper; Rosenblat 2018), and the introduction of robots into pharmacies (Barrett et al. 2012) and hospital settings (Beane 2018; Beane and Orlikowski 2015). For instance, Christin (2017) compared and contrasted the use of predictive algorithms in two occupational domains where elements of work could be automated or supplemented by data mining capabilities and computational power, first for assessing criminal profiles in criminal justice and second in predicting article readership in online journalism. This literature demonstrates professionals’ the variety of responses to the introduction of new technology in their workplace.

These studies demonstrate how it is not enough to show was algorithms are, but how they are used in practice: “A practice only exists as it is materialized in specific times, places, texts, artifacts, bodies, infrastructure, and so on” (Beane and Orlikowski 2015, p. 1553). Doing so involves accounting for the range of technological features in addition to social and organizational features (Zammuto, Griffith, Majchrzak, Dougherty, & Faraj, 2007). As transformative technological change unfolds, symbolic dimensions of working with AI tools are especially important, as individuals make sense of AI over time (Prasad 1993; Weick 1990). Reflecting the evocative nature of AI that holds multiple meanings for different individuals, reflecting their fears, desires, concerns, and needs. This study of AI in radiology seeks to explain what AI and its technological features represents to radiologists, and how those representations shape and are shaped the diagnostic decision-making process.

High Stakes Knowledge Work and Artificial Intelligence

While organizational research on automating or augmenting work with AI is currently in its early days, specific research in economics seeks to characterize and measure the anticipated impact of technological advances like AI on future labor markets at the level of tasks, jobs, and overall labor markets. Using tasks characteristics, Dhar (2016) argues that tasks with a higher degree of predictability and a low cost of making an erroneous judgement are more likely to be subject to AI automation than less predictable tasks with higher costs. Furthermore, Frey and Osborne (2017) conclude that just under half of jobs in the United States are highly susceptible to automation based on the degree of perception and manual manipulation, creative intelligence, and social intelligence required to a job's core tasks. It is essential to extend scholarly discussion beyond predictions and averages to understand what is happening on the ground in work contexts where AI has already begun to augment or automate tasks. This study addresses this need, examining diagnosis work in radiology, a high-stakes task that requires synthesis and expertise.

At its core, making diagnostic decisions is a form of knowledge work, which has grown as a focus of organizational research in recent decades (Barley and Kunda 2001; Bechky 2006; Boland and Tenkasi 1995). Knowledge is conceptualized as an ongoing process embedded in individuals' everyday work practices (Nicolini 2012; Orlikowski 2002) that emerges from its social and organizational context (Lave and Wenger 1991; Suchman 1987). Researchers have been studying how knowledge work is impacted and reconfigured with the introduction of new technology, which is among the set of social and organizational factors that facilitate and shape knowledge work (Knorr-Cetina 1999; Latour 1987; Orlikowski 2002; Zammuto et al. 2007). For instance, scholars analyze how the rise of digitalization provides tools and methods, facilitating communication, and expanding work roles and relations (Boland et al. 2007).

However, it is unclear what effects AI will have on processes of knowledge production and evaluation. Individuals within a given field share practices about what type of knowledge is valid and how to engage with information, which are increasingly being challenges or reinterpreted in light of new advanced technology. It is critical for studies to closely investigate contexts in which these tools are already in use. Doing so will allow deeper understanding of how AI is embedded within and shaping such processes, going beyond characterizations of such tools as either useful or not. As new technologies are adopted and used, there is potential for such tools to fundamentally challenge or disrupt epistemic cultures and professionals' existing ways of knowing (Kaplan 2011). For instance, Dougherty and Dunne (2011) show how the use of digital tools and methods led to the creation of new kinds of knowledge in drug discovery. In the context of financial trading, Knorr Cetina (2016) traces the disruption of knowledge assumptions as trading evolved towards relying on algorithmic models rather than human experts. Pachidi and colleagues (2019) trace the struggles and ultimately elimination of a sales team asked to use a predictive tool that challenges their long-standing practices of knowledge production and evaluation. It is unclear what will be the role of humans in knowledge processes where technology is seemingly capable of "generating" novel insights on its own as technology detects patterns and relationships in datasets (Christin 2017; Faraj et al. 2018).

Methods

Research setting

This research is based on an in-depth qualitative field study of a Radiology Department at Urbanside Hospital[1], a tertiary medical center in a major United States city. Diagnostic radiology is a medical specialty in which medical imaging technology is used to help diagnose and treat a wide range of diseases. If a physician suspects a diagnosis for a current patient, they may refer the patient to have medical imaging, which was analyzed by diagnostic radiologists. Depending on the patient scenario, radiologists may be involved in defining imaging parameters or participating in the image acquisition process. The radiologist opens the patient's complete study, black and white images, patient history information and reporting software fill four large monitors. At Urbanside, diagnostic radiologists spent anywhere between a few

[1] All names are pseudonyms.

minutes to an hour analyzing one case, depending on a number of factors including imaging modality (e.g., X-ray, CT scan, MRI, ultrasound, mammogram), the number and complexity of suspected ailments, the patient's medical history, number of prior images available, and so forth. After analyzing the case (described in detail in the first section of the findings), the radiologist described their overall impressions and diagnostic conclusions in a detailed report that is sent back to the referring clinician.

Data collection

Starting in fall of 2018, I immersed myself in the field of radiology, attending professional conferences, symposia, trainings, and vendor presentations, to understand the challenges and opportunities on the field's horizon from the radiologist perspective. Ethnographic field work began in January of 2019 and spanned three sections of radiology at Urbanside that were actively using AI tools in their section: Chest radiology, Breast imaging, and Pediatrics radiology. A total of 34 radiologists were observed in this study, including 22 attendings, 4 fellows, and 8 residents.

Observation

I spent between three and four days a week observing radiologists' day-to-day work (Van Maanen 1988). This involved documenting detailed observations of the radiologists' diagnostic decision making process in real time in a notebook, which were supplemented upon leaving the hospital each day. Because Urbanside radiologists trained medical students, residents, and fellows, it was possible to capture radiologists' detailed explanations of their decision making process and diagnostic reasoning. In total, I observed over five hundred diagnostic decisions. I paid particularly close attention to the way radiologists used technological artifacts and how they spoke of these artifacts in practice. Approximately one third of the observed diagnosis decisions involved using AI tools. During these instance, it was specifically documented how the tool was used, to what extent radiologists' considered the AI output, and any verbal or nonverbal reactions made by the radiologist and among radiologists.

Interviews

Data was also collected through a total of 27 semi-structured interviews (Spradley 1979). Fourteen informal interviews took place as radiologists conducted their work or during short breaks between tasks, covering topics such as why they selected their particular sub-specialty or what types of cases they prefer. Thirteen formal interviews allowed me to deepen our understanding of the interviewees' perspective of what it means to be a radiologist, their diagnostic decision making process, and the role of various technologies in their work. All formal interviews and some informal interviews are recorded (with informants' permission and consent) and transcribed.

Documentation and artifact

Finally, I collected documentation and artifact data. First, I captured technical artifacts of the AI tools: pieces of code, accuracy outputs of the AI models, training data sets, FDA submissions, published technical research, and detailed documentation materials. I also captured artifacts produced and used by the radiologists: medical notes taken, photographs and drawings of images on their monitors, technology training materials, and departmental communications. These artifacts were used to deepen my understanding of radiologists' work and the diagnostic decision making process.

Data analysis

In keeping with the principles of grounded theory development, I engaged in regular analysis during and throughout the observational period (Charmaz 2014; Glaser and Strauss 1967). In the early stages, I conducted open coding to capture a broad range of emerging themes, including "seeking colleague's input", "searching for more information", "showing frustration over the image quality", "expressing the need to move quickly," and so forth. Following grounded theory methods, I constantly compared the emerging themes and categories across ongoing data collection. Within the first few months, the extreme time pressure of the radiologists' work and their frequent expressions of doubt and question asking while making diagnoses were striking. I was also surprised to observe such variation in how AI tools were used across the range of cases.

Following these emerging themes, I conducted additional targeted data collection and analysis focusing on the diagnostic decision making processes for the specific categories of diseases where AI tools were implemented: routine breast cancer screening (Breast section), routine lung nodule search (Chest section), and bone age assessment (Pediatrics). Through this phase, additional themes emerged related to how radiologists interacted (or not) with the tool, including "questioning AI validity," "disagreeing with the tool assessment," "confirming original diagnosis," and "expressing doubt". I observed multiple patterns of how and when individuals evoked AI tools and to what extent these assessments were incorporated into user' final determinations.

To understand this variation, I analyzed additional data related to their implementation and use. I studied the differences in the technical features and their respective historical patterns of use. Analysis was conducted to understand how these differences were related to how variation in how the radiologists perceived and used the tools in practice. I also conducted additional rounds of comparative analysis that led to deeper understanding of how radiologists with different levels of expertise approached the task and use of certain AI tools differently.

Once I established a grounded understanding of the diagnostic processes involving AI, I turned to the theoretical literature of advanced technology in knowledge production and evaluation processes. I iterated between analyzing empirical observations and relating them to prior literature to further enhance the formal theory development (Glaser and Strauss 1967). Through this stage, I found the role of ambiguity to be particularly important, and thus I continued to analyze the data taking these concepts and theoretical relationships into account, which is described in the next section.

Findings

1.0 Making high stakes diagnostic decisions at Urbanside

Making accurate and complete diagnoses[2] involved high stakes decision making that had the potential to dramatically impact patients' lives as well as radiologists' professional reputations. Urbanside radiologists were cognizant that their decisions had immense consequences for their patients. They remarked on the stakes for patients as they read cases, such as when Hattie (Breast) reflected on patients' heightened level of anxiety as they waited for their cancer screening results: "Half of them are out of their mind when they come in here. They're all in this anxiety, and it's just a whole thing. They're all nervous. They all come here and think they have cancer." The impact of their diagnostic determination weighed considerably on Urbanside radiologists while they performed their work. One afternoon, Leslie (Chest) reported a severe diagnosis and released a deep exhale as she picked up the phone to relay the results to the patient's referring physician: "It's going to ruin her day. It's going to ruin everything for her."

Making accurate and complete diagnostic decisions was also consequential for radiologists' professionally. Urbanside radiologists' awareness of one another's strengths and weaknesses factored into their daily interactions, especially regarding who was and was not consulted for help. This was exemplified by Lola (Breast), who explained why her confidence in her diagnosis increased after consulting a colleague with a reputation for making excellent calls: "I was looking for someone whose opinion I would trust on a tricky case. Different people have different strengths. Her (the doctor she consulted) recall rate is 6% (far better than the department average)! There's a few people who have both time and practice, and a really good eye. They're people who we would like to read more like." Radiologists made references to what they perceived to be their strengths and their desires to develop their weaknesses. It was especially common for junior radiologists to discuss their goals and efforts towards improving their decisiveness. In the following exchange, Raj, a resident in the Breast section, hesitated while making a diagnostic decision, explaining how he was "super edgy on these cases," or struggling to select among diagnosis alternatives. Shaw, who

[2] When we refer to the "diagnostic decision", we refer to the set of findings and impressions radiologists provided to referring clinicians in their radiology report. The report may include one or more possible diagnosis "differentials" and may explicitly rule others out. Upon received the report, clinician then take further steps towards confirming the diagnosis, which is typically beyond the scope of radiologists' work and this study.

just began his one-year fellowship[3] in the Breast section, was reading alongside him: “Me too. But I’ve really got to change. That’s what this year is for.”

Diagnostic errors were associated with major professional consequences for radiologists’ reputation, heightening the stakes of each decision they made. Radiologists bore sole responsibility for their evaluations of patients’ cases, and their signed reports are permanently archived in patients’ medical records. Most past reports resulted in little to no fall-out, but on occasion, diagnosis errors came to light that called the radiologist’s ability and reputation into question. In some instances, radiologists discovered one another’s errors by finding disease in a patient’s current study and noticing earlier signs of the disease were detectable in prior imaging – earlier signs that a previous radiologist failed to detect. The traceable nature of the archived radiology reports and the implications for their reputations contributed to radiologists’ heightened level of concentration: “Since we have 20 years’ worth of mammograms, you can always go back and say, ‘Oh, they missed it.’ So we have to be really careful,” (Irene, Breast). Urbanside radiologists had formal and informal practices for notifying the responsible radiologist of the errors they detected. One morning in the Chest reading room, a fellow found signs of a severe condition in a patient’s current imaging. When she realized a senior attending had not commented on visible signs of the disease in a study earlier that year, she consulted another attending about how to address the situation, “Do you think we should tell the person who read the original chest radiograph?...I would want to know if I missed it.” The attending replied, “We always want to know...When this happens, we do let one another know. I’ll email him. Otherwise how will we ever get better?”

Diagnosis errors could lead to medical malpractice lawsuits that jeopardized radiologists’ reputation as well as their financial and job security. The possibility of being sued was not absent from radiologists’ daily activities. This was the case for Leroy (Chest), who specifically included a carefully worded phrase in each radiology report he signed. He explained how he began using this sentence after a challenging lawsuit he was involved in years ago: “I got sued a long time ago for a study I read correctly. The patient died from something related to the heart, and the lawyer kept saying, ‘Why didn’t you talk about her cardiomyopathy on her report?’ And I kept saying, ‘This is not the right test for that diagnosis.’...So I don’t know. This (the sentence he always inserted) might not protect me from a lawsuit, but it makes me feel better.”

Struggling to make accurate and complete diagnostic decisions

There were two distinct dimensions that led to diagnostic errors: the possibility of misinterpreting the study and the possibility of missing important details. Urbanside radiologists, like Nadia (Pediatrics) frequently voiced concerns that stemmed from one or both of these possibilities: “You don’t want to get a case wrong. You don’t want to make the wrong diagnosis, obviously, or miss something.”

Radiologists were concerned that the way they interpreted a study could lead to diagnostic errors. The primary source of diagnostic evidence for radiologists was patients’ medical imaging. Accurately interpreting the subtleties of medical imaging required expertise, which radiologists’ developed through extensive training and apprenticeship: “Radiology residency is 5 years for a very good reason. It’s just, the image looks like different shades of grey! That’s what I thought when I was first a medical student...Someone once called it ‘retinol mileage’. You just have to see a lot,” (Irene, Breast).

Yet even with ample “retinol mileage”, the ambiguity introduced by medical imaging created regular opportunities for misinterpretation and error. High degrees of professional judgement and subjectivity were involved in translating visual signals into a variety of different possible diagnoses. Complex anatomy and diseases varied widely within patient populations and elevated the possibility for misinterpretation: “For breast imaging, everyone’s breast looks different. Like, a liver is a liver, a gull bladder is a gull bladder. Depending on your age, menopausal status, your weight, additional risk factors, there is huge variation. It really requires a lot of concentration for us to look for cancer,” (Irene, Breast). Radiologists frequently expressed their continued deliberation and second-guessing of cases they completed earlier in the day: “I don’t know, I really hope we’re right about that,” (Rita, Pediatrics).

[3] Radiology training typically involves completing medical school, five years of radiology residency, at least one year of fellowship, and all required professional examinations prior to becoming an attending radiologist.

Urbanside radiologists were on edge when faced with the possibility that they made a diagnostic error on a past patient. One afternoon, the head of the Pediatrics section called the central phone in the Pediatrics reading room. Lori answered the call, listened, and then confirmed the spelling of the patient's name out loud while looking up the record. Across the room, Shira sat up straight in her chair, recognizing the name of a patient she read a few hours before. She listened to Lori's end of the conversation on full alert and probed her for more information: "Did I do something wrong? (No response from Lori) Why are we looking at this case? Did I screw up? (No response) Why is she asking us to look it up? (No response) This makes me very stressed."

Given the professional judgement and subjectivity involved, diagnostic errors were ultimately made and caused radiologists angst and strife. During a particularly challenging moment, a radiology Chest fellow broke down in the reading room, "There are so many things going on but not everything fits. I feel like I've been doing everything wrong lately....I'm just muddling my way through things case after case. Sometimes you take cases and feel right on. Other days you are just drowning," (Sylvie). In response, Leslie (Chest) offered encouragement to the distressed fellow, explaining how errors are difficult to avoid when making interpretations: "All of us have been presented with a bad outcome. We all have misses. Interpretation is hard. It's not necessarily like you weren't looking or paying attention. It's like, you interpreted it, right? And we can be right or wrong when we interpret things," (Leslie, Chest). Many radiologists echoed the reasoning that making mistakes is inherent when making subjective diagnostic decisions: "There's always going to be stuff you don't know. There's always things that come up that I've never seen before...Everyone has misses. That's part of the job is just being able to accept that," (Emily, Breast).

Regarding the second dimension, Urbanside radiologists were also acutely aware of the possibility of missing consequential details. Consuming and integrating every relevant detail in the imaging and patient history required unyielding attentiveness. In addition to a patient's vast medical history, each study comprised multiple high resolution images or sequences of images (50-70 consecutive slices). Each available image also included multiple alternative views and numerous options and features for highlighting or subduing various anatomical aspects: "There is so much volume of data on the films to deal with," (June, Chest).

Even with a focused and methodical approach, relevant details were sometimes obscured by the complicated nature of human anatomy and disease. In the case of detecting breast cancer on mammogram images, radiologists rigorously searched the patient's white-appearing tissue for miniscule dots that were also white-appearing: "It's like looking for a snowball in a snowstorm," (Caroline, Breast). In the case of searching for lung nodules in a lung CT study, Leslie (Chest) struggled while trying to visualize an area of tissue nestled behind a rib: "Am I hallucinating a nodule? It's behind two ribs, so it's impossible to get a good look. Ah I think it's a vessel. Yeah I don't think it's a nodule. Windowing really helps. Ah yeah, I'm pretty sure." By "windowing", or lowering the brightness of bone, she could better visualize the hidden area and concluded it was an innocuous blood vessel instead of a nodule (which would have suggested cancer).

The possibility of missing details was exacerbated by certain limitations of imaging techniques. Radiologists commonly ordered additional medical imaging if they suspected the additional views could help them visualize the anatomy more completely. In the following instance, Lola (Breast) wanted classify a suspicious area deep in the patient's breast as benign, but hesitated due to her inability to see that tissue perfectly (areas deep in the breast are difficult to capture using mammogram): "It's (the suspicious tissue) probably down here (just outside the image border). It's probably related to her surgical scar. It's probably in this patch of tissue somewhere. If I could see clearly in this area, I wouldn't be so concerned." Despite the many logical reasons Lola listed in support of making a benign classification, she ordered additional images to be certain she was considering every possible consequential detail.

Practicing doubt during diagnostic decision making

Urbanside radiologists commonly exercised doubt practices to help them cope with the subjectivity and complexity involved in diagnostic decision making. Exercising doubt involved seeking out evidence that would contradict or support tentative diagnosis theories. Doubt practices involved radiologists asking themselves questions, seeking colleagues' opinions, and acquiring additional imaging or patient information. These practices helped radiologists prevent premature conclusions, thoughtfully weigh conflicting information, and thoroughly consider every available detail. In this instance, June (Chest) questioned her tentative diagnosis that a specific nodule was stable (benign) and searched for additional evidence to either prove or disprove her theory: “Is that it? I guess so. I think it’s there. I want to be more convinced.” She scrutinized images in prior studies until finally confirming that the nodule was benign, as it had a stable size and shape over several years. In the following case, Leroy (Chest) gestured to a thick haziness in the patient’s lungs as he sat back in his chair and posed questions to himself that helped him consider a range of diagnosis alternatives. “I’m trying to picture: what’s going on here? Could it just be mild edema? Yes, it could be. But could it also be something else? Like, if you did a CT would you see little nodules and airspace? I think it could be, yes. So I think it could be bronchopneumonia. And I’m going to put possible drug-reaction too because that can give you this kind of appearance.”

Soliciting one another’s input was another way radiologists exercised doubt. Once a radiologist made a tentative assessment, they oftentimes asked a colleague to weigh in, either over the phone, via IM, or in person. This was particularly common when assessing potentially life-threatening diagnoses, as Shira (Pediatrics) described, “We do a double and triple check when we read malignancies. Nobody wants to miss a tumor. Nobody wants to miss a reoccurrence of cancer.” In the following instance, Kyle (Breast) asked his colleague Lola (Breast) to weigh in on a mass that was catching his eye, and Lola conveyed her opinion that the mass was benign. Kyle asked how she reached that conclusion, and Lola directed him to “look at (the image from) 2012. Zoom in. Really zoom in. It’s (the mass) circumscribed, and it’s uniform. Do you agree with me?” Kyle hesitated due to how prominent the mass looked in the current 2019 image. Lola offered additional explanation into her reasoning, “I didn’t make the mass go away, but I established stability. It’s got a uniform shape (in the 2019 image) and in 2012 it’s got a uniform shape.” Nodding, Kyle scrolled through the full set of images again and agreed that the mass was benign. A formalized version of this practice is referred to as double-reading, whereby two radiologists independently interpreted a case and converged their assessments into a single report. Yet, due to the time-consuming and resource-intensive nature of double-reading, it was not widely instituted at Urbanside.

Temporal demands on diagnostic decision making

At Urbanside, time was incredibly precious, as radiologists were under immense pressure to achieve rapid turnaround times for high volumes of cases. In a particularly revealing interaction, the head of the Breast section explained her tendency of referring to the Breast imaging section as “the gulag”: “Welcome to the gulag! That’s what I call it, the gulag….Where they work you to death and they throw maybe a cup of water at you, maybe a slice of bread every other day. That’s what the gulag is. It’s a horrible, horrible place.” It was not uncommon to observe Urbanside radiologists working continuously at their workstation for hours without leaving their desk or checking their phone or email. The authors of a recent commentary published in prominent academic Radiology journal also emphasized the intense and demanding culture of modern radiology environments[4]: “Time is a scarce resource in the modern fast-paced, high-pressure clinical [environment]. Radiologist fatigue has been a long-standing concern but is being exacerbated by expanding practices, growing examination volume, and increasing complexity of imaging data that need to be interpreted.”

Time pressure further exacerbated Urbanside radiologists’ likelihood to make incomplete or misinterpreted interpretations that could result in diagnostic errors. As they strained to keep up with the incoming flow of studies throughout the day, their ability to provide diligent evaluations diminished: “At the end of the day, there’s like a mad rush to try to get through all the cases. And you’re most tired at the end of the day. You’re trying to plow through, trying to increase your speed and get through the material,” (Gianna, Chest). As Gianna suggested, radiologists recognized the potential consequences of moving too rapidly through their work: “To me, you know, there’s too much human error. People are already going so fast,” (Nadia, Pediatrics).

Thus, Urbanside radiologists faced a deep tension between the temporal demands of their organizational context and their dedication to making accurate and complete diagnostic decisions. Sitting at her workstation at one morning, Irene (Breast) was concerned by the queue of cases which was already overwhelming and continued to grow: “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,” (Irene, Breast). Consequently,

[4] Hsu, W. & Hoyt, A. 2019. Using Time as a Measure of Impact for AI Systems: Implications in Breast Screening. *Radiology: Artificial Intelligence,* 1 (4): 1.

Urbanside radiologists were focused by reading efficiently and avoiding wasted time and effort. Every doubt-related practice Urbanside radiologists enacted incurred a potential costs, in terms of time, within the high-pressure environment: “If I was really looking, like taking my time, teaching medical students, I think I would have seen it. But I think if I was just trying to get through as many studies as possible, I'm not so sur,” (Sylvie, Chest).

Making accurate and complete diagnostic decisions was both extremely high stakes and extremely prone to error. They worked diligently by exercising doubt practices to avoid making harmful errors. Intense temporal demands not only intensified the likelihood of making such errors, but also amplified the costs for exercising doubt. Building upon this understanding, the following section presents how the use of AI tools related to Urbanside radiologists’ diagnostic practices and struggles.

2.0 Coping with AI-induced-ambiguity during diagnostic decision making

Urbanside radiologists widely shared expectations that AI tools were capable of enhancing their daily work. In particular, they pointed to AI as a means to improve the speed at which they could perform time-consuming routine tasks: “It’s kind of menial labor at this point, just to be using calipers to measure whether a nodule is bigger or smaller. It’s something that should be automated in the future, and that will make a huge difference in the speed we can get through cases,” (Brady, Chest). They often described the alternative tasks they could spend their time performing if AI would alleviate the tedious aspects of their work. In the following case, Shira (Pediatrics) expressed that if AI could do tedious measuring tasks for scoliosis patients, she could focus her expertise on higher-risk anatomy (such as the lungs and abdomen): “We want it (AI) for scoliosis….It’s incredibly tedious. It’s not interesting work…If something (AI) could do that for me, then I could look at the lungs and look at the abdomen myself. That’s where my eyes are going to be the most important.”

Using AI for the bone age assessment task

Radiologists in the Urbanside Pediatrics section expressed especially high expectations that AI could help them perform “bone age assessments”. Assessing bone age[5] involved determining how a patient’s skeletal maturity related to their chronological age by comparing a patient’s hand x-ray to an atlas of standard images of children’s hand at each consecutive year of life. To make these assessments, a volumous text book was situated in the radiologist’s lap, and she looked back and forth comparing different age standards to the appearance of the current patient’s x-ray on their monitor. Excitement grew around the idea of implementing a bone age AI tool, as Nathan underlined: “I think AI 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.” So Urbanside Pediatrics radiologists implemented an AI tool (developed by an external research institution) that automatically generated a bone age assessment in the background: radiologists made an independent tentative diagnosis then viewed a numerical bone age generated by the AI which was pre-populated into their radiology report.

Although they expected AI to improve their speed and diagnostic outcomes, the Pediatrics radiologists’ actual use of the tool in fact impeded those outcomes. In the occurrences where the assessments converged (about a third of the time), the radiologist expressed happy surprise and signed the report: “I would give him 12. So what did the computer give him? It gave 12 years 2 months! So that’s close enough. I gave 12, and it gave 12 years 2 months. How it knew that? I have no idea,” (Nathan). However, in the vast majority of cases, the radiologists’ tentative diagnosis conflicted with the AI assessment[6]. Dealing with the conflicting information required radiologists to spend more time completing the previously routine task, as illustrated by the following case. Nadia’s eyes flicked back and forth between the standard images and the patient’s x-

[5] Urbanside Pediatrics radiologists typically conducted less than five bone age assessments per day. Referring clinicians ordered a bone age study to determine the extent to which a patient could tolerate growth hormones. For instance, it would be ill-advised to use such treatment for patients’ whose skeletal bone age is more advanced than their chronological age.

[6] It is difficult to say for certain whether the radiologist or the AI tool is “more accurate”. However, it is common practice for the radiologist’s (or a consensus of radiologists’) opinion to be considered the gold standard.

ray as she formed her independent diagnosis: “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 same. I think he’s between 8 and 9. The machine says between 9 and 10. Closer to 10 actually!” Responding to the AI’s assessment, Nadia cocked her head to the side in frustration at the discrepancy: “Now I’m going to try to find why it said that.” She consulted the text book again, flipping to the 10 year standard, and shook her head, “but I feel he’s not that close to that (10 years). I think the machine’s overestimating. To me, it’s 8 and 9. Sometimes I change my mind after hearing what the machine says. But this time, no. No.”

In cases of divergence, ambiguity surged due to the conflicting information. This was the case for Nadia, who struggled in the face of the AI-induced ambiguity: “Sometimes it (the AI) would give me bone ages that would make me re-think what I said and go, ‘Ok, maybe.’ And I would adjust closer to it (the AI assessment). But sometimes, I think, ‘This is way off.’ So I don’t know. I just don’t know,” (Nadia). In the face of the ambiguity of conflicting assessments, radiologists began doubting the diagnosis and had to answer the question of what was the most accurate diagnosis?

Radiologists coped with increased ambiguity by questioning the AI’s validity as well as questioning their own diagnosis. Regarding their doubt about the AI’s validity, the professionals raised questions about the tool’s development, diagnosis process, and quality relative to their own. They questioned how the tool was trained, what data was included, and who were the radiologists who provided the training assessments: “If I knew it was all done by Pediatric radiologists who had read thousands and thousands of bone ages, I think I would consider it probably more vetted, more trust-worthy,” (Lori). They also asked questions about how the tool’s diagnostic process related to the way they were professional trained to make such diagnoses: “I feel sometimes the algorithm is looking at different things than I am…We are trained to use the phalanges as the gold standard, but there’s also carpal bones and the radius ulna…We don’t look at those as much. I don’t know how much weight the AI gives to that,” (Rita). Finally, questioned were raised as to whether the AI could provide high quality assessments than their won: “I wanted to see how well the computer could do. Because I believe I’m an expert, so I wanted to see how good this computer is,” (Nathan).

At the same time, radiologists were also questioning their own diagnosis. Upon viewing conflicting AI assessments, the ambiguity led them to second-guess their own expertise for this task: “I’m often disagreeing with the model, and maybe it’s just me and I don’t know how to read bone ages,” (Nadia). To resolve discordant assessments, radiologists re-evaluated the standard image at the bone age suggested by the AI tool. They reasoned about what logic the AI may be drawing upon and whether it differed from their own professional approach. In the following instance, Nathan began his evaluation at the 18 year old standard in the text book, which was the patient’s chronological age: “You see here (on the patient’s x-ray), all fissured. See here in the book, there’s still a tiny physis here, you can still see a light line here.” A bright white line (the “tiny physis”) ran horizontally between the dark grey bones in the 18-year-old standard image in the text book; no such horizontal white line was visible on the patient’s x-ray (which Nathan described as “all fissured”). This suggested to Nathan that the patient’s bone age is more advanced than 18 years, since the bones had grown even closer together than the 18-year old standard. Upon reading the AI assessment of 17 years 4 months, Nathan sighed heavily and returned to the text book: “Look here (pointing to the 18- year-old standard), it’s open. And here (pointing to the patient’s x-ray), it’s not open at all. So I’m going to give it 19. Either way, it is still read as normal. But still, I disagree with the computer on this one.”

Ultimately the use of the bone age assessment AI tool induced additional ambiguity which Urbanside radiologists struggled to reduce using doubting and questioning practices. The following section examines two other AI tools used by radiologists in the Chest and Breast sections.

Using AI for routine lung and breast cancer screening tasks

A common task in the Breast and Chest sections is performing routine cancer screenings. This task is highly subject to making errors due to the possibility of radiologists missing consequential details in the complex studies. To screen a patient for cancer, radiologists carefully scrutinized either series of images in CT studies (when screening for lung cancer) or highly-magnified mammogram images (when screening for breast cancer). They searched for tiny abnormalities throughout the images that could be the beginning stages of life-threatening diseases. In the following case, June (Chest) described the difficulty of searching for lung nodules, as she scrolled vigorously through the full length of the CT series of images for several minutes: “In this view, the lungs are squished together and there are some hidden areas…I can’t do both lungs at the same time. I go through this section, then this section there, then that area. I don’t see anything major jumping out. Hopefully I'm not missing anything." Radiologists described the unwavering attention and special expertise required for cancer screening tasks to avoid missing a potentially life-threatening finding: "There's different things were looking for in the mammogram (image): masses, calcifications, skin thickening, changes to the tissue, axillary lymph nodes, distortion. All these different things to see if there's disease or not," (Kyle, Breast).

Urbanside radiologists in the Breast and Chest sections have been using AI tools during cancer screening tasks for several years. Radiologists refer to these tools as "CAD" tools (computer-aided detection tools). CAD tools were provided by external vendors and each one was developed, trained, and validated uniquely for the anatomical area and distinct nature of the task (e.g., searching for nodules in lung tissue vs searching for masses and tissue distortion in breast tissue). Therefore, the technical structure of the underlying CAD tools differed. However both tools signaled potential abnormalities to the radiologists by presenting a series of circles, triangles, and stars as an overlay on the patient's image: "This star is saying that there's a density here and this triangle says there are calcifications that are suspicious here," (Sophie, Breast).

When radiologists viewed the results presented by the CAD tools, they experienced increased ambiguity. They typically did not use CAD to guide their initial decision making but instead evoked the tool (and its associated ambiguity) as the last step in their diagnostic process. Using in this manner provided a way to exercise doubt against their own tentative assessment, as June (Chest) explained, "I only turn on the CAD at the very end. If I had it on early, as I'm scrolling through, my eyes just pop straight to what they marked. So I like to look through all of it on my own first and see everything I can then turn it on afterwards." Due to their focus on making complete and accurate diagnosis decisions, radiologists regarded CAD as a version of a second opinion or double-read for the high stakes cancer screening tasks: "A little assist is not the worst idea…If we're not going to double-read mammos (using two radiology professionals) then something like an AI double-read can be helpful," (Lola, Breast).

In contrast to Pediatrics radiologists who questioned both themselves and the tool, Chest and Breast radiologists coped with AI-induced ambiguity by only mildly questioning themselves. Using CAD led radiologists to briefly pause and reconsider their assessment, but rarely did it result in meaningful diagnostic changes. Radiologists responded in this manner when dealing with ambiguity stemming from two paths of conflicting information. In the first path, CAD flagged areas the radiologist did not initially detect or find concerning: "You go look for nodules to mark them, and you use the CAD and find any nodules that you missed," (Brady, Chest). In these cases, radiologists mildly questioned themselves by quickly assessing each flagged area. However, they rarely altered their decision because they concluded the markings were false positives: "We discount a lot of the stuff it picks up because it's just picking up nothingness," (Kyle, Breast). In the second path, ambiguity emerged when radiologists flagged areas of concern that CAD did not detect, such as in the following case: "CAD didn't mark anything on this one. It didn't even mark the thing that caught the radiologists' attention!" (Lola, Breast). In these cases, radiologists did not waver in their determination that the area was concerning, and they concluded the non-marked areas were false negatives: "They (CAD markings) are not the primary determinant…If there's something that's concerning to you, based on your initial interpretation, that the CAD is saying, 'Oh, this looks normal,' that's not something you would use. You wouldn't use that information and say, 'Oh, we're not going to biopsy it,'" (Sophie, Breast).

For radiologists in these sections, there was no questioning of CAD's validity or of the tools' diagnostic processes. Years of concluding that CAD's markings were frequently false positives or false negatives had crystalized radiologists' perceptions of the shortcomings of the tools. This was particularly the case for the CAD tool used in Breast imaging: "CAD sucks. That's why nobody likes it." (Lola, Breast). CAD often flagged findings that were obvious to expertly trained Breast radiologists, but did not typically suggest additional findings that the radiologist missed: "CAD increases sensitivity but not specificity….If you follow-up on everything it said, your specificity would be down in the basement. Sensitivity is over the top, and specificity is way under," (Hattie, Breast). As Hattie explained, in addition to CAD reporting too many false positives ("sensitivity is over the top"), it failed to strengthen radiologists' weaknesses and provided too many false negatives ("specificity is way under"). These views were stabilized over years of interacting with CAD, so we did not observe radiologists engage in question these AI tools in practice.

3.0 Shifting routine decision making tasks to nonroutine

Even for highly routine and by-the-books tasks like bone age assessment and routine cancer screening, Urbanside radiologists use of AI tools in these settings varied widely. In all three sections, the use of AI tools induced additional ambiguity that required radiologists to take action to resolve. Depending on the user and their context, there were varying degrees of questioning and scrutinizing. This was exacerbated by the ongoing tension between the perceived usefulness of using AI tools and the perceived costs of doing so, which resulted in further variation in how and when different radiologists used the tools in practice. Finally, ongoing use of the bone age AI tool ultimately led to variation in how Pediatrics radiologists read bone age studies, as some individuals changed their practices to mimic the AI tool. This variation is described in the remainder of this section.

Trading offs perceived costs and usefulness of AI

Across all three sections, Urbanside radiologists' own professional judgement consistently outweighed the AI assessment. The previous section revealed how Chest and Breast radiologists frequently overruled CAD by disregarding its common false positives and negatives. In the case of the bone age task, Pediatrics radiologists corrected the AI assessment in over 68% of the cases[7], suggesting their perception that the tool was inferior to their professional ability. This practice was surprising given that Urbanside radiologists commonly expressed how low they held the bar of expertise required to make routine and by-the-book bone age assessments: "Quite frankly, I think anyone can do a bone age…It's pattern recognition. Right. I don't think it's a very sophisticated thing," (Lori). Such perspectives were widely shared in the section: "Do you have a high school education? You can read a bone age," (Miguel), and, "There's nothing scientific about this. It's just, ya know, kind of comparison." (Rita).

However, just as Urbanside radiologists did not robotically accept the tools' assessments, they also did not automatically record their own assessment. There were occasions when the tools provided critical insights: "I always check CAD, but I take it with a grain of salt basically because it picks up things that are normal, that are totally negative. But once in a while it definitely picks up things that you looked at yourself and you totally ignored, that you just couldn't see," (Shirma, Chest). At the same time, constant supervision was required to use the tool effectively: "CAD is often wrong…You do have to quality check it (CAD) a lot," (Brady, Chest) and, "I still have to just double-check it, and then make sure, and then change it," (Miguel, Pediatrics). Thus, Urbanside radiologists worked to vigilantly monitor the tools in order to avoid becoming complacent, as explained by Nathan (Pediatrics): "If people become complacent and they feel like, 'Oh, this thing (AI) is great! This can do everything!' Then I think it can have a real downside."

Consequently, the time radiologists spent deliberating cases while using AI actually increased. Radiologists engaged in questioning the AI and more thoroughly securitized their own diagnosis. These practices required valuable time that radiologists generally did not have, as Lola (Breast) expressed, "A lot of the stuff CAD draws attention to, you wish it didn't." Time was spent coping with ambiguity and reconciling conflicting information: "It slows me down. Because I see what it says, and I don't trust it," (Nadia, Pediatrics) and, "It's actually just another step for me to do. Like, how many times did I click that thing on and off just to evaluate what it's picking up?" (Kyle, Breast). At the same time, the additional time was rationalized by some radiologists in favor of helping improve their diagnostic outcomes: "You're willing to sacrifice the fact that you might pick up false stuff because you really are trying to catch the things that are most concerning," (Raj, Breast).

Tradeoffs between the perceived usefulness of using AI and its perceived costs resulted in variation in how and when these tools were used. In the case of the CAD tools, senior radiologists expressed more critical attitudes towards CAD based on their years of experience overruling CAD's false positives and negatives: "I mean, the cancers? I know what it looks like, ya know? I don't need you (CAD) to tell me what cancer looks like. Clearly benign things? Also, I don't need you," (a senior Breast radiologist). Therefore, some senior radiologists used CAD less consistently and with less attentiveness: "I sometimes hit it (CAD) at the end. But like, when I'm in a rush, I don't always hit it. It's probably not appropriate, but it's just reality," (a senior Breast radiologist). In contrast, radiologists in training at Urbanside (fellows and residents) had

[7] Adjustments made to the assessment varied from one month to two and a half years.

accumulated less first-hand exposure to the specific benefits and limitations of CAD. They also tended to compensate their less-developed professional expertise by consulting CAD more consistently and earnestly: "For me, I kind of suck at finding lung nodules just by scrolling through like this. So I know this (CAD) is going to help them stand out a little better," (Cal, Chest Fellow). Consequently, these more junior individuals were particularly appreciative of the stronger peace of mind they experienced after consulting the tool: "You want to have confidence in your decision. It's nice to have help…especially on those question mark cases," (Kyle, Breast Fellow).

Changing standardized practices to mimic AI

In the case of the bone age task, there was increased deviation in how Pediatrics radiologists performed this previously routine task. This deviation is related to the introduction and ongoing use of the bone age AI tool. Prior to the implementation of the tool, Pediatrics radiologists at Urbanside followed a standard practice when evaluating bone ages: use the whole-numbers associated with the standards included in the professionally-instituted *Greulich and Pyle*[8] atlas (e.g., reporting a bone age of "6 years" or "between 14 and 15 years"). Miguel described his adherence to this standard practice even when the patient's x-ray appeared somewhere between the standard years: "Normally, in the past, if I thought the bone age was something in between…I would say in my report – and this is just because somebody taught me this – I would say, 'The bone age is in between 12 and 13 years.'"

However, in the reports generated by the bone age AI tool, a more granular bone age was used which included a numerical value for the years and the months (e.g. reporting a bone age of "5 years 2 months"). This approach was in direct conflict to radiologists' typical training, as explained by Rita, "This AI program kind of interpolates what it thinks it's closest to. It doesn't always give you the exact number. It gives you a kind of interpolation between standards. So we don't typically do that."

Accordingly, some radiologists balked at the AI's divergence from their standard approach. When Shira read bone ages, she remained aligned with the standard approach she was trained to use and pointed to her confusion of the tool's approach: "It is confusing to me. It would give me a bone age of like, 10 years 1 month. My human brain does not know what 10 years and 1 month looks like. I know what 10 years looks like. I know what 11 years looks like…Even saying something is at a half way point between those two, we don't know! That doesn't make any sense to me." Others agreed and continued to report using their professional standard, overruling any assessment whereby the tool reported a month-level granularity: "It will give you things like 11 years 8 months. Like how do you get that? I can't break something up into – I would maybe be able to break it up into quarters, but not like *twelfths*. So I don't know what to do when it gives me like 8 months. I'm like, 'Uh-uh. No.'" (Nadia).

Alternatively, some radiologists observed the AI reporting at a higher granularity and accepted this as a valid way to read bone age studies. This acceptance was granted despite the fact that it deviated from their previous standards, as explained by Nathan, "If they (the AI) gives a weird bone age, like 14 years 2 months, and I think it's 14 years – that's close enough to me! Given that the standard deviation is 28 months, I'm okay with reporting using the two months." This acceptance was so strong that some radiologists changed their own practices and began independently reading bone ages in a way that mimicked the AI's granularity. One afternoon in the Pediatrics reading room, Rita dictated her decision of the current bone age case: "13 years and 4 months." The author asked how did arrive at that decision, and she explained: "It's a little bit less than the standard 13 years 6 months. (She paused and shrugged) Not everybody does that. I guess I've been doing that since the AI started doing it."

In each of the three sections, Urbanside radiologists performed previously routine decision making tasks with higher degrees of variation upon using AI tools. This was unexpected, given their expectations that AI would simply perform the tedious task and improve radiologists' efficiency. Not only did it increase variation, but the use of AI tools induced additional ambiguity into the previously routine tasks. Thus, as radiologists worked to cope with the ambiguity, their efficiency actually diminished.

[8] *Greulich and Pyle's Radiographic Atlas of Skeletal Development of the Hand and Wrist*. 1959. Stanford University Press.

Discussion

This study provides one of the first studies of the use of AI for conducting professional decision making tasks. It illuminates the important link between ambiguity and the use of such tools in high judgement decision making activities. Weick (1990) conceptualized early decision support technologies as "equivoque", or open to multiple interpretations and subject to uncertainty. As technologies become more advanced, he argued it would be essential for users to be able "to reason about the deep structure of new systems" (Weick, 1990: 39). Indeed, tools using AI methods are increasingly opaque and inscrutable, whereby layers of complex reasoning takes place out of sight and inaccessible to end-users (Faraj et al. 2018). Even prior to the current breakthroughs in AI technology, scholars have been focusing on the extent to which knowledge work professionals have been able to understand and critically evaluate assumptions underlying technologies they encounter at work (Anthony 2018; Bailey and Barley 2010; Dodgson et al. 2007; Knorr-Cetina 1999).

This study illuminates how issues of opacity are particularly salient and consequential in decision making tasks that involve intelligent technologies like AI tools. Researchers have been examining how work practices are impacted when the assumptions embedded in technology were not immediately obvious to users (Bailey et al. 2012; Boland et al. 2007). For tasks involving professional decision making, the process of logical reasoning that an AI tool (or a human) used to reach a given conclusion is highly relevant. Particularly in healthcare contexts, where professionals bear legal, professional, and moral responsibility for their decision making, they were highly unsatisfied by the lack of information available to them about the tool. As prior literature predicts, these professionals engaged in limited degrees of questioning practices (Anthony 2018). However, the technical opacity and the intense time pressure precluded them from engaging in extensive search or prolonged questioning. Lacking information intensified the degree of ambiguity they faced and was critical to their ultimate integration (or lack of integration) of the tools' assessments, as exemplified by this Pediatrics radiologist, "Knowing more about how the algorithm was trained and using what standard would certainly change my confidence. Because I would have some understanding of it. It might also decrease my confidence. So either further decrease it or increase it. It certainly would. Of course," (Shira). Although they were unable to peer inside the technical composition of the AI tools, they made comparisons between the AI output and other available information to form their assessment of AI validity.

Opacity is also critical to how AI tools are perceptions and use became stabilize over time, as ongoing, situated practices accumulated. One of the hallmark features of modern AI tools is their ability to continually learn and dynamically improve their accuracy (Faraj et al. 2018). However, we find that users' perceptions crystallized early on, despite the fact that the quality of the tool could significantly improve over time. Since the tools operate behind the scenes, outside of the realm of users' visibility, professionals were unable to interrogate the tool or have any sense of the tools' changing quality. Ultimately this study expands our understanding of the relationship between intelligent tools' characteristics related to opacity and the degree of ambiguity in decision making processes. This study shows that the resulting ambiguity led actors' to generally overrule opaquely-generated outputs and rely on their own known professional judgement process, despite the underlying quality of the AI tool.

Finally, this study reveals the unexpected dynamics and resulting ambiguity involved in implementing AI into high judgement work. As technical advances of AI continue to explode across industries and in healthcare in particular, the wave of "digital exceptionalism" is also rapidly growing (Lancet 2018). Digital exceptionalism describes the view of many futurists, technology optimists, advocates, developers, and vendors who evangelize the remarkable outcomes that can result from the widespread adoption and use of advanced innovations in intelligent technologies. However, there is an equally prominent set of voices advocating for caution and the need to gather grounded understandings of how such tools are poised to impact medical practices, professionals, and patient outcomes in the coming years and decades (Marabelli and Newell 2019; Recht and Bryan 2017). This study provided such in-depth examination. Doing so revealed the complexities of implementing and using AI tools in clinical settings, where, rather than alleviating uncertainty, AI introduced additional ambiguity into medical decision making. Moreover, the implementation of AI did not result in a tidy story of work automation or complete resistance, but instead, the use of such tools led to new configurations of professionals and AI working in partnership to make decisions with life or death consequences.

References

Amabile, T. 2019. "GUIDEPOST: Creativity, Artificial Intelligence, and a World of Surprises Guidepost Letter for Academy of Management Discoveries," *Academy of Management Discoveries*. (https://doi.org/10.5465/amd.2019.0075).

Anthony, C. 2018. "To Question or Accept? How Status Differences Influence Responses to New Epistemic Technologies in Knowledge Work," *Academy of Management Review* (43:4), pp. 661–679. (https://doi.org/10.5465/amr.2016.0334).

Bailey, D. E., and Barley, S. R. 2010. "Teaching-Learning Ecologies: Mapping the Environment to Structure Through Action," *Organization Science* (22:1), pp. 262–285. (https://doi.org/10.1287/orsc.1090.0511).

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

Barley, S., and Kunda, G. 2001. "Bringing Work Back In," *Organization Science* (12:1), pp. 76–95. (https://doi.org/10.1287/orsc.12.1.76.10122).

Barrett, M., Oborn, E., Orlikowski, W. J., and Yates, J. 2012. "Reconfiguring Boundary Relations: Robotic Innovations in Pharmacy Work," *Organization Science* (23:5), pp. 1448–1466. (https://doi.org/10.1287/orsc.1100.0639).

Beane, M. 2018. "Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail," *Administrative Science Quarterly*. (https://doi.org/10.1177/0001839217751692).

Beane, M., and Orlikowski, W. J. 2015. "What Difference Does a Robot Make? The Material Enactment of Distributed Coordination," *Organization Science* (26:6), pp. 1553–1573. (https://doi.org/10.1287/orsc.2015.1004).

Bechky, B. A. 2006. "Talking About Machines, Thick Description, and Knowledge Work," *Organization Studies* (27:12), pp. 1757–1768. (https://doi.org/10.1177/0170840606071894).

Boland, R. J., and Tenkasi, R. V. 1995. "Perspective Making and Perspective Taking in Communities of Knowing," *Organization Science* (6:4), pp. 350–372. (https://doi.org/10.1287/orsc.6.4.350).

Boland, R., Lyytinen, K., and Yoo, Y. 2007. "Wakes of Innovation in Project Networks: The Case of Digital 3-D Representations in Architecture, Engineering, and Construction," *Organization Science* (18:4), pp. 631–647. (https://doi.org/10.1287/orsc.1070.0304).

Cameron, L. Working Paper. *Making Out While Driving: Control, Coordination, and Its Consequences for Algorithmic Labor*.

Charmaz, K. 2014. *Constructing Grounded Theory*, Thousand Oaks, CA: Sage.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165–1188.

Christin, A. 2017. "Algorithms in Practice: Comparing Web Journalism and Criminal Justice," *Big Data & Society* (4:2). (https://doi.org/10.1177/2053951717718855).

Davenport, T. H., and Kirby, J. 2016. *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines*, New York, NY: HarperBusiness.

Dhar, V. 2016. "When to Trust Robots with Decisions, and When Not To," *Harvard Business Review*.

Dodgson, M., Gann, D. M., and Salter, A. 2007. "'In Case of Fire, Please Use the Elevator': Simulation Technology and Organization in Fire Engineering," *Organization Science* (18:5), pp. 849–864. (https://doi.org/10.1287/orsc.1070.0287).

Dougherty, D., and Dunne, D. D. 2012. "Digital Science and Knowledge Boundaries in Complex Innovation," *Organization Science* (22:5). (https://pubsonline.informs.org/doi/abs/10.1287/orsc.1110.0700).

Erickson, I., Robert, L., and Nickerson, J. 2018. "Workshop: Work in the Age of Intelligent Machines," in *GROUP '18 Proceedings of the 2018 ACM Conference on Supporting Groupwork Pages 359-361*.

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), pp. 62–70. (https://doi.org/10.1016/j.infoandorg.2018.02.005).

Frey, C. B., and Osborne, M. A. 2017. "The Future of Employment: How Susceptible Are Jobs to Computerisation?," *Technological Forecasting and Social Change* (114), pp. 254–280. (https://doi.org/10.1016/j.techfore.2016.08.019).

Glaser, B., and Strauss, A. 1967. *Discovering Grounded Theory*, Chicago, IL: Aldine Publishing Company.

Kaplan, S. 2011. "Strategy and PowerPoint: An Inquiry into the Epistemic Culture and Machinery of Strategy Making," *Organization Science* (22:2), pp. 320–346. (https://doi.org/10.1287/orsc.1100.0531).

Kasparov, G. 2017. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, PublicAffairs.

Kellogg, K. C., Orlikowski, W. J., and Yates, J. 2006. "Life in the Trading Zone: Structuring Coordination Across Boundaries in Postbureaucratic Organizations," *Organization Science* (17:1), pp. 22–44. (https://doi.org/10.1287/orsc.1050.0157).

Knorr-Cetina, K. 1999. *Epistemic Cultures. How the Sciences Make Knowledge*, Cambridge, MA: Harvard University Press.

Knorr-Cetina, K. 2016. "What If the Screens Went Black? The Coming of Software Agents," in *Beyond Interpretivism? New Encounters with Technology and Organizations*, Berlin, Germany: Springer.

von Krogh, G. 2018. "Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing," *Academy of Management Discoveries* (4:4), pp. 404–409. (https://doi.org/10.5465/amd.2018.0084).

Lancet. 2018. "Is Digital Medicine Different?," *The Lancet* (392:10142), p. 2694. (https://doi.org/10.1016/S0140-6736(18)31562-9).

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, MA: Harvard University Press.

Lave, J., and Wenger, E. 1991. *Situated Learning: Legitimate Peripheral Participation*, Cambridge University Press.

Lycett, M. 2013. "'Datafication': Making Sense of (Big) Data in a Complex World," *European Journal of Information Systems* (22:4), pp. 381–386. (https://doi.org/10.1057/ejis.2013.10).

Marabelli, M., and Newell, S. 2019. *Reflections on Algorithmic Decision-Making in the Healthcare Industry*, , April 2.

Nicolini, D. 2012. *Practice Theory, Work, and Organization: An Introduction*, Oxford University Press.

Orlikowski, W. J. 2002. "Knowing in Practice: Enacting a Collective Capability in Distributed Organizing," *Organization Science* (13:3), pp. 249–273. (https://doi.org/10.1287/orsc.13.3.249.2776).

Orlikowski, W. J. 2010. "The Sociomateriality of Organisational Life: Considering Technology in Management Research," *Cambridge Journal of Economics* (34:1), pp. 125–141. (https://doi.org/10.1093/cje/bep058).

Orlikowski, W., and Scott, S. 2014. "What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector," *Organization Science* (25:3), pp. 868–891.

Pachidi, S., Berends, H., Faraj, S., and Huysman, M. 2019. "Symbolic Actions During The Introduction Of A New Technology: The Ironic Change Of A Regime Of Knowing," *Organization Science*.

Prasad, P. 1993. "Symbolic Processes in the Implementation of Technological Change: A Symbolic Interactionist Study of Work Computerization," *Academy of Management Journal* (36:6), pp. 1400–1429. (https://doi.org/10.5465/256817).

Recht, M., and Bryan, R. N. 2017. "Artificial Intelligence: Threat or Boon to Radiologists?," *Journal of the American College of Radiology* (14:11), pp. 1476–1480. (https://doi.org/10.1016/j.jacr.2017.07.007).

Rosenblat, A. 2018. *Uberland: How Algorithms Are Rewriting the Rules of Work*, Univ of California Press.

Simon, H. A. 1987. "Making Management Decisions: The Role of Intuition and Emotion," *The Academy of Management Executive (1987-1989)* (1:1), pp. 57–64.

Spradley, J. 1979. *The Ethnographic Interview*, New York, NY: Holt, Rinehart and Winston.

Suchman, L. 1987. *Plans and Situated Actions*, Cambridge, U.K.: University of Cambridge Press.

Valentine, M., Retelny, D., Rahmati, N., Doshi, T., and Bernstein, M. 2017. "Flash Organizations: Crowdsourcing Complex Work by Structuring Crowds as Organizations," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.

Van Maanen, J. 1988. *Tales of the Field: On Writing Ethnography, Second Edition*, Chicago, IL: University of Chicago Press.

Weick, K. E. 1990. "Technology as Equivoque: Sensemaking in New Technologies," in *Technology and Organizations*, The Jossey-Bass Management Series, San Francisco, CA, US: Jossey-Bass, pp. 1–44.

Zammuto, R. F., Griffith, T. L., Majchrzak, A., Dougherty, D. J., and Faraj, S. 2007. "Information Technology and the Changing Fabric of Organization," *Organization Science* (18:5), pp. 749–762. (https://doi.org/10.1287/orsc.1070.0307).