




Friend or Foe? Teaming Between Artificial Intelligence and Workers with Variation in Experience

Weiguang Wang,^a Guodong (Gordon) Gao,^{b,*} Ritu Agarwal^b

^aSimon Business School, University of Rochester, Rochester, New York 14627; ^bCarey Business School, Johns Hopkins University, Baltimore, Maryland 21202

*Corresponding author

Contact: wwang90@simon.rochester.edu,  <https://orcid.org/0000-0001-7568-7685> (WW); gordon.gao@jhu.edu,  <https://orcid.org/0000-0002-2336-9682> (G(G)); ritu.agarwal@jhu.edu,  <https://orcid.org/0000-0002-8907-5652> (RA)

Received: August 23, 2021

Revised: April 23, 2022; December 6, 2022;
March 20, 2023

Accepted: May 3, 2023

Published Online in Articles in Advance:
October 11, 2023

<https://doi.org/10.1287/mnsc.2021.00588>

Copyright: © 2023 INFORMS

Abstract. As artificial intelligence (AI) applications become more pervasive, it is critical to understand how knowledge workers with different levels and types of experience can team with AI for productivity gains. We focus on the influence of two major types of human work experience (narrow experience based on the specific task volume and broad experience based on seniority) on the human-AI team dynamics. We developed an AI solution for medical chart coding in a publicly traded company and conducted a field study among the knowledge workers. Based on a detailed analysis performed at the medical chart level, we find evidence that AI benefits workers with greater task-based experience, but senior workers gain less from AI than their junior colleagues. Further investigation reveals that the relatively lower productivity lift from AI is not a result of seniority per se but lower trust in AI, likely triggered by the senior workers' broader job responsibilities. This study provides new empirical insights into the differential roles of worker experience in the collaborative dynamics between AI and knowledge workers, which have important societal and business implications.

History: Accepted by Kartik Hosanagar, information systems.

Funding: This work was supported by Inovalon [Sponsor of the Health Insights AI Laboratory].

Supplemental Material: The data files and online appendix are available at <https://doi.org/10.1287/mnsc.2021.00588>.

Keywords: artificial intelligence • human-AI teaming • worker experience • productivity • healthcare • medical coding

1. Introduction

In recent years, the resurgence of artificial intelligence (AI) is capturing substantial attention in the business world (Manyika et al. 2017, PwC 2017). Unlike past generations of AI that required humans to translate knowledge into rules, a distinctive feature of contemporary AI is its ability to learn from data (Brynjolfsson et al. 2018a, Hosanagar 2019). Driven by machine learning (ML)-based approaches, contemporary AI has redefined the power of machines and represents a dramatic paradigm shift compared with previous generations of the technology (Taddy 2018).

Riding on this new power, AI that uses modern developments in machine learning (and its subfield of deep learning) is rapidly encroaching on the domain of high-level cognitive tasks that used to be reserved for humans (He et al. 2015, Mnih et al. 2015). As noted by Korinek and Stiglitz (2017, p. 2), “it is clear that [AI] has the potential to disrupt labor markets in a major way, even in the short and medium run, affecting workers across many professions and skill levels.” A recent paper (Brynjolfsson 2022) suggests that, rather than focusing on how

machines might replace humans, it is critical to direct research attention to how AI can augment humans.

Unsurprisingly, in the foreseeable future, the question of how human workers can best team with AI presents a significant challenge. An emerging stream of literature has begun to examine teaming intelligence (i.e., how humans and AI can work together) (Johnson and Vera 2019, Bogert et al. 2021). Although most studies focus on how AI can be optimally designed to better team with human workers (Chakraborti et al. 2017, Clement et al. 2020, Zhang et al. 2020, Ahn et al. 2021, Lebovitz et al. 2021, Siu et al. 2021), limited research seeks to understand the influence of factors on the human side, a notable gap that we aim to address.

Our study focuses on a critical aspect of human capital—worker experience. Evidence suggests that workers' experience could profoundly affect their relationship with AI, especially in knowledge-intensive jobs that are the focus of heated debate about how AI is affecting the value of human workers' experience (Swap and Leonard 2014, Henry-Nickie 2017). To illustrate, AI can help detect tumors in medical images with high accuracy (McKinney

et al. 2020), suggesting that it could potentially make a radiologist's experience less valuable. Alternatively, there are reasons to believe that AI can complement human workers' experience. If AI largely automates a subset of well-defined tasks, workers with rich experience would be more productive in leveraging it, as it would free up cognitive resources to focus on the tacit and complex aspects of their work (Pakdemirli 2019, Raisch and Krakowski 2021).

In light of these controversies, recent empirical work has started focusing on how work experience affects human-AI teaming. We contribute to this emerging discourse by examining the interplay of contemporary AI with workers of different experience types and levels. Our empirical analysis is conducted at the level of a specific task, allowing us to shed light on individual worker factors that affect the benefits of AI in knowledge work. Findings from this study have important implications to help executives and policy makers build a more productive relationship between the two forms of intelligence.

For this investigation, we implement a machine learning-based AI in a knowledge work setting in a publicly traded U.S. company in the healthcare sector. Our AI is built for medical coding and identifies patient conditions in medical charts, a critical task in medical reimbursement. This use of AI for a medical coding task offers an ideal setting for studying the role of experience in human-AI work teaming for several reasons. First, the specific medical coding task is an archetypal complex and cognitively nonroutine task. Thus, findings from this study may be generalizable to similar knowledge-intensive jobs. Second, the AI in this study is an exemplar of a typical machine learning-based AI that is being adopted in business. Given that most machine learning models focus on well-defined tasks, the mainstream use of AI in business is to augment rather than replace humans as previously discussed. Consistent with current industry practice, the AI in this study is designed to augment human intelligence. Third, medical coding has well-defined output with strict quality controls, thereby offering appropriate metrics for a detailed empirical analysis of productivity gains. Finally, our study setting is such that we can rule out resistance to learning new technologies, a common confounding factor with experience. As such, our study offers nuanced insights on how experienced workers team up with AI differently than those with less experience.

Using detailed data from the coding of 1,120,831 patient charts over a one-year period, we measure the differential impact of AI on productivity conditional on workers' experience levels. We further conduct a series of qualitative analyses and a laboratory experiment to gain a deeper understanding of the nuances and underlying dynamics of AI's impact. We examine two critical facets of knowledge worker experience. In the human capital literature, *amount* and *time* are widely recognized

as the two most widely used quantitative measures of experience (Quinones et al. 1995, Tesluk and Jacobs 1998). Based on that, we employ the two dimensions of experience: the volume of a worker's tasks (McDaniel et al. 1988, Ford et al. 1991) and the worker's tenure on the job (Lance et al. 1989, Vance et al. 1989). Volume-based experience or task experience indicates practice in the specific task (volume of medical charts in our context). This measure is well aligned with AI's expertise. It allows us to examine whether the knowledge of AI and humans is complementary or substitutable.

The second measure of experience, tenure on the job, examines how a worker's organizational experience interacts with AI. Recently, scholars have suggested that successful teaming with AI requires effective collaboration at the organizational level (Agrawal et al. 2021, Bresnahan 2021). This time-based experience or seniority experience incorporates understanding of not only the specific task but the entire job. Senior workers accumulate organizational experiences over time, which may influence their beliefs, attitudes, and behaviors in the context of the human-AI interaction (Wagner et al. 1987, Judge 1994). This measure, therefore, captures factors beyond task-specific knowledge.

Our empirical analysis yields several novel findings. First, we identify an interesting pattern in the human-AI teaming dynamic; AI boosts productivity for all workers but more so for workers with greater task-based experience and less so for workers with greater time-based experience. A follow-up qualitative analysis and a laboratory experiment unveil the reason why senior workers did not benefit from AI as much as their junior colleagues; their wider scope of work and greater concern for the welfare of the organization induce a higher degree of sensitivity to AI's imperfections, which in turn, erodes trust and causes pushback.

This study makes significant contributions to AI research examining field implementations of AI at its current technological stage. We add to a nascent stream of literature that examines the teaming of humans and AI, shedding light on the complexity of this relationship from the perspective of variation in workers' experience. Our study advances understanding of interactions between human experience and AI by extending human experience to a multidimensional measure, where task experience enhances human-AI teaming, whereas seniority experience creates resistance. To the degree that AI is predicted to transform work and society, our study offers practical guidance on how to amplify AI's potential. We establish how AI performs with different levels of human capital, and although our research setting is healthcare, the findings are generalizable to knowledge-intensive work in other domains. As such, they yield important managerial implications for the wider utilization of AI in business and help to show how human and machine intelligence can work together.

2. Theory and Background

2.1. Contemporary AI

Artificial intelligence is shaping our lives dramatically and is now broadly heralded as a potential stimulus for an economic revolution (Brynjolfsson et al. 2018a, Hosanagar 2019). Industry projections estimate that AI's contribution to the U.S. economy will be \$15.7 trillion by 2030, constituting a boost of up to 26% in gross domestic product (GDP) (PwC 2017). AI has evolved over the past few decades, transforming from early rule-based systems to ML algorithms that have the ability to learn. Rule-based systems require explicit acquisition and codification of human knowledge, whereas ML-based AI learns directly from data. Given its lack of reliance on explicit human knowledge, contemporary AI using recent developments in ML is fundamentally different¹ from previous generations of AI (Brynjolfsson et al. 2019); it exhibits both unique advantages and shortcomings.

One advantage is that ML algorithms enable AI to learn with high efficiency. Equivalent knowledge that human workers may have spent years accumulating can be rapidly extracted from data in hours. Moreover, domain experts no longer need to explicitly code their knowledge into rules for AI, eliminating what used to be an effort-intensive and expensive task of knowledge acquisition (Hayes-Roth et al. 1983). Rather, machine learning algorithms detect patterns from data by iteratively and recursively exploring complex connections between inputs and outputs to establish the “relationship” (Jordan and Mitchell 2015). The derived “relationship” is able to reflect complex connections, supporting more accurate predictions than previous generations of AI.

Despite the advantages, contemporary AI's learning ability also has some shortcomings. First, current artificial narrow intelligence, characterized by its limited focus (Jajal 2018, Fjelland 2020, Zysman and Nitzberg 2020), learns from data with a clearly defined objective function to optimize. As a result, it is highly concentrated on this goal, paying little attention to other aspects. Because AI is completely reliant on the training data and predefined learning mechanisms (Lum 2017, McCraden et al. 2020), its performance will be confined to the training data set. Any contextual or environmental information is invisible to the AI if it is not already present in the preselected data or in the case of reinforcement learning, the preselected environment.

Second, the complex relationships discovered by ML are significantly harder to interpret than explicit rules used by previous generations of AI. Often, this opacity creates uneasiness in accepting AI's decisions (Burt 2019), thus negatively affecting the human-AI team relationship. This is similar to a human learner; it is hard to accept a person's judgment if the person does not explain the rationale underlying the decision. The lack of interpretability often results in high-performing AI not being

implemented (Siu et al. 2021). For example, in the finance industry, analysts are often required to use traditional AI models instead of contemporary ones because the business must explain how decisions are made (Sarkar 2018). The inability to fully interpret the reasoning process of AI at its current stage (Knight 2017) frequently degrades users' trust and willingness to team with AI. This is also the driving force behind the emerging efforts in explainable AI development (Ribeiro et al. 2016).

Finally, the absence of common sense in the AI's model is a major limitation (McCarthy 2007, Zysman and Nitzberg 2020) to effective human-AI teaming. Humans accumulate a rich and wide-ranging set of experiences in life, spanning a gamut from basic math, understanding emotion, the concept of holidays, social structures and family relationships, etc. AI is not equipped with such common sense because of the specificity of its training, and current AI can be easily confused by subtle distinctions that are effortlessly resolved by humans and especially so for linguistic tasks (Davis and Marcus 2015, Tandon et al. 2018, Huang et al. 2019). For example, the lack of a sense of temporal progression (a patient cannot progress from age 40 to age 30) and social structure (a patient can only have one biological father but more than one sibling) can reduce a medical AI's performance when it uses related information (age and family history) to predict the patient's condition.

Because of the distinctive nature of contemporary AI, findings from previous generations of expert decisions support systems (many of which are rule based) might not be readily generalizable to this new generation of technology. ML-based AI's unique features call for new studies to shed light on its interactions with human work experience.

2.2. Human-AI Teaming

Human-AI teaming has become an important emerging topic in various research fields to accelerate the applications of AI into organizational practice. One active research area focuses on how to train AI for better teaming. In this stream, Nikolaidis and Shah (2013) propose the “human-robot cross-training” framework to allow AI and human workers to learn iteratively. Other work includes that by Carroll et al. (2019), who develop an adaptive AI agent to accommodate human behaviors, and Sadigh et al. (2018), who allow the AI underlying autonomous cars to incorporate other human drivers' potential responses for optimized coordination. Methodologies are also being developed for continual training or updating of existing AI for better teaming with humans (Bansal et al. 2019) and for helping AI better understand human agents' goals (Puig et al. 2020).

Similarly, in the information systems field, various aspects of AI system design are being examined, including the role of transparency and interpretability of algorithms (Clement et al. 2020, Ahn et al. 2021, Siu et al.

2021), types of tasks that should be delegated to AI (Fügenger et al. 2022, Dai and Singh 2023), and how to evaluate AI performance (Lebovitz et al. 2021). Given that a key feature of contemporary AI is its learning ability, studies are also exploring the role of human knowledge in complementing AI (Fügenger et al. 2021) and how AI affects organizational learning systems (Sturm et al. 2021).

We note that limited attention has been focused on heterogeneity among users as a potential determinant of the human-AI teaming efficacy. From a reputation perspective, Dai and Singh (2020) provide one of the first theoretical models of the level of physician expertise in AI adoption. Among the few empirical studies that address variation across individuals, experience has been identified as an important factor affecting the outcomes of human-AI teaming (Ge et al. 2021). In a study conducted in a simulated laboratory setting using a general framework based on metacognition, Jussupow et al. (2021) show that experienced physicians are more likely to reject AI's input. In a study with a similar, albeit more limited focus than ours, Allen and Choudhury (2022) conjecture that human workers with more domain experience can better leverage AI, but they also tend to exhibit greater algorithm aversion.

Building on and extending prior research, in this study we theoretically and empirically distinguish task-specific domain experience from broad domain experience and examine their differential roles in the human-AI teaming relationship.

2.3. Worker Experience

How does worker experience moderate the effects of human-AI teaming on productivity? Studies from the history of technological innovation give cause for concern that AI may disproportionately advantage certain types of human workers (David 2015, Decker et al. 2017, Wu and Kane 2021). Such heterogeneity in human collaboration with new technologies manifests through the pathway of technology-skill complementarity (Goldin and Katz 1998). In recent decades, the changes in information technology (IT) have been argued to be an especially powerful instance of skill-biased technology change (SBTC) (Krusell et al. 2000). Using data from the 1990s, Bresnahan et al. (2002) find a strong tendency for IT to favor skilled labor. If AI continues the trend of SBTC, we expect that AI would favor workers with greater experience.

Measures for worker experience have been developed in different domains and with different levels of specificity (Ostroff and Ford 1989, Klein et al. 1994, Tesluk and Jacobs 1998). The most widely used experience measures are (1) the length of time in a specific position (Borman et al. 1993) or a job (McDaniel et al. 1988, McEnrue 1988) and (2) the volume of a task (Lance et al. 1989, Vance et al. 1989). Quinones et al. (1995, pp. 891–892) divide the experience measure into three dimensions: *amount*, *time*, and *type*. Amount measures “refer to numerical counts

such as the number of times a task was performed or the number of different jobs held in an organization,” whereas time measures reflect tenure, such as months or years in the job. Although the amount dimension might appear to overlap with the time measure, Ford et al. (1991) show dramatic differences in the challenge and complexity of task assignments for individuals with equivalent jobs or organizational tenure.² The difference between time and amount dimensions is also reinforced by other research (Tesluk and Jacobs 1998, Ng and Feldman 2010). The *type* measure is usually context specific and hard to objectively measure; therefore, it is categorized as a qualitative indicator (Tesluk and Jacobs 1998).

In their studies of AI adoption, Agrawal et al. (2021) and Bresnahan (2021) emphasize that the factors influencing AI adoption exist at the organizational level and are not limited to the individual task. Therefore, we study human-AI teaming by examining workers' experience for both the individual task (task experience) and related to the organization (seniority experience). We define amount-based experience, or task experience, as the volume of the specific task. Learning from repeated task execution, workers update their knowledge through each interaction. The updated knowledge can further improve the worker's ability to predict the effectiveness of different strategies in subsequent steps (Dunlosky and Hertzog 2000). Therefore, increased familiarity with the task is expected to promote task performance of individual workers (Schmidt et al. 1986, Goodman and Leyden 1991), and studies find that greater task experience is associated with higher performance (Maruthappu et al. 2015).

In contrast, time-based experience, or seniority experience, is a “broad” measure of work experience based on years in the job and reflects the influences a worker is subject to with the passage of time. In our context, this time-based experience is measured using the number of years in the job (i.e., organizational tenure) (Ng and Feldman 2010). With a greater length of time in the job, workers not only absorb more knowledge regarding the focal tasks and surrounding workflows, but they are also exposed to organizational influences to a greater extent. As documented in the literature, person-organization fit and goal alignment can be increased by high seniority experience (Wagner et al. 1987, Judge 1994). Workers with longer standing within the organization are more likely to be concerned with responsibility, power, and authority. Influences from the social environment and the organization can further affect attitudes and behaviors, including loyalty to the company, job security, esteem, and risk preferences (Mortimer and Lorence 1979, Veiga 1981, Kohn and Schooler 1982, Slocum et al. 1985, McCauley et al. 1994).

Multidimensional measures of work experience are critical for advancing understanding of human-AI interactions. Human work experience is traditionally considered homogeneous in the AI literature (Mnih et al. 2015,

Dong et al. 2016). Allen and Choudhury (2022) explained their findings by proposing the existence of two countervailing forces underlying human work experience: ability and aversion. However, both forces were associated with a unidimensional work experience measure because of data limitations in their study. Given the complexity of human work experience, the origins of knowledge workers' ability and aversion are better disentangled using distinct experience measures.

2.4. Interplay Between AI and Human Workers

2.4.1. AI and Task-Based Experience. As noted, contemporary AI is characterized by its narrow focus (Fjelland 2020). If an AI is successfully trained on a task-specific data set, AI can substitute for a worker's task experience. Because human task experience is accumulated through task completion, which is recorded in the training data, plausibly, a human worker's task experience is substitutable by AI. To illustrate, in medicine, AI outperformed typical medical staff in making diagnoses based on medical images (Gulshan et al. 2016). This means that a radiologist's experience in reading images is less valuable. In this case, workers with less experience could benefit more from AI. In addition, AI can reduce well-documented errors in human judgment (De Martino et al. 2006, Danziger et al. 2011), such as anchoring (Tversky and Kahneman 1974) and recency effects (Tzeng 1973). Given the significance of human judgment in the economy (Kahneman and Tversky 1979, Kahneman 2011), AI can further help less experienced workers by addressing potential bias in human cognition and assuring quality output.³

On the other hand, the substantial difference in the learning mechanisms of humans and of AI might make the human experience less substitutable. The training process of contemporary AI that focuses on the objective function can lead to the acquisition of unique, idiosyncratic knowledge as compared with human workers. AI may observe subtle patterns in the training data and take shortcuts to achieve high performance (LeCun et al. 2015). For example, models often resort to shortcuts by picking up unexpected patterns from data to make accurate predictions. A clinical study (Jabbour et al. 2020) discovered that the high level of performance observed from image reading AI models was a result of the AI inferring a disproportionately high connection between having a pacemaker and congestive heart failure. Although AI's learning incorporates the identification of more subtle correlations, as the authors demonstrated, they are not universally applicable, which leads to an immediate failure if tested on a different group of instances. Indeed, such failures in learning are part of why contemporary AI is labeled as "narrow" (Fjelland 2020).

By contrast, rather than accumulating experience through back propagation, human knowledge workers update their "decision function," not only driven by the final outcome but through a broad range of intermediate

knowledge (Rudin 2017). Human learning also involves explicit long-term memories that are lacking in current AI (Welleck et al. 2018, Dinan et al. 2020), rendering the human learning curve substantially different from AI's learning process. These differences in learning create an opportunity for human workers to complement AI using their breadth of experience. Indeed, a popular use case of contemporary AI is to assist humans making decisions and providing intermediate input rather than final decisions. In order to assess the quality and veracity of AI outputs, the human user needs to possess a breadth of knowledge, such as knowledge of practice, data-related knowledge, and exposure to the data. Interpretation of AI's output involves understanding its real meaning and comprehending the reasoning of AI. This logic suggests that workers with in-depth knowledge and proficiency around their work might be able to make fuller use of AI.

Furthermore, as noted, the lack of common sense is a major limitation of contemporary AI (McCarthy 2007, Zysman and Nitzberg 2020). Human decision makers, however, are equipped with common sense, increasing the divergence between human task experience and AI. To the extent that the task knowledge of the human worker is different from that of the AI, the two sets of knowledge should complement each other. Therefore, we hypothesize that given the same specific task, human workers with greater task experience benefit more from AI than those with less task experience.

2.4.2. AI and Seniority Experience. In contrast to task experience, seniority experience reflects a greater breadth of knowledge, facilitating a deeper comprehension of the task at hand and affording senior workers more advantages. Because current AI is so focused on a narrow task and is blind to all other experiences, such as the upstream and downstream activities of the specific task, senior workers' broader experience can enhance AI's performance by incorporating more unseen but related knowledge. Moreover, seniority experience reflects an understanding of the entire organization. To the extent that the implementation of AI is more of an organization-wide change instead of being restricted to the individual task level (Agrawal et al. 2021, Bresnahan 2021), seniority experience plays a critical role for workers' teaming with AI.

However, from an alternative perspective, it is plausible that senior workers could be disadvantaged in teaming with AI because of organizational and social factors. As knowledge workers stay longer in an organization, they cognitively adjust their perceptions. These factors might be critical to a human worker's attitude toward new technologies and may, therefore, affect how the worker interacts with the AI. Existing studies have also documented the possible negative impact of work experience on the utilization of new technologies (Tan and Netessine 2020). In an AI implementation among sales agents, Luo et al. (2021) also identified the lower benefit

for the top-ranked workers than their middle-ranked colleagues. AI, as a highly intelligent but early-stage technology, is likely to be more vulnerable to the negative factors.

Trust has been identified as a critical factor that explains human resistance to technology (Li et al. 2008, McKnight et al. 2011). Research has examined trust through three dimensions: ability, integrity, and benevolence (Mayer et al. 1995, McKnight et al. 2002). Prior studies are predominantly focused on technologies related to communication and information processing that serve as facilitators to provide necessary information to users, with high-level decision making remaining vested in the human. Compared with previously studied technologies, several unique characteristics of contemporary AI have amplified the critical role of trust (Rossi 2018, AI HLEG 2019).

First, AI's abilities may be challenged more by senior workers than their junior colleagues. It is well known that experts are more likely to display a resistance to judgments that are not their own (Liu et al. 2017). The self-confidence engendered by virtue of their experience could lead them to discount the recommendations of another (Bradley 1981). This is especially the case with the contemporary AI that does not offer perfect performance (Dietvorst et al. 2015). Studies have documented users' resistance to AI because of its neglect of "uniqueness," which in this case, means that AI may fail to capture a patient's idiosyncratic characteristics or circumstances (Longoni et al. 2019). Compared with novice users, senior workers are more likely to spot imperfections in AI, triggering stronger doubt in its intelligence (Bansal et al. 2019) and lower trust in the AI's output. Because senior workers have greater responsibility for overall organizational performance, their attention may be disproportionately attracted to the few errors made by AI.

Second, senior workers may experience discomfort when AI's integrity cannot be determined because of its opacity, and as a result, they may resist accepting it. The integrity of any entity that is inherently difficult to understand can be challenging to ascertain. Senior workers' concern about losing control may be exacerbated by the fact that many algorithms employed by advanced AI are notorious for their lack of interpretability (Pasquale 2015, Bhatt et al. 2020, Baniecki et al. 2021). By virtue of the length of their experience, senior workers develop well-defined reasoning for executing tasks. Because AI's reasoning is not visible, the lack of transparency can degrade the perception of AI's integrity and therefore, may decrease senior workers' trust in AI. It is interesting to note that in a recent study, Ahn et al. (2021) show an innovative way to increase user trust in AI by providing performance feedback, suggesting subtle dynamics between ability and integrity.

Third, senior workers may be more concerned with the benevolence of AI because of their high-level perspectives. Following the stream of organizational studies,

employees with long organizational tenure increase the alignment between their personal goals and the organizational goals (Wagner et al. 1987, Judge 1994). Whenever AI makes moves that senior workers cannot comprehend, they may worry that it could potentially jeopardize their careers or adversely affect the organization. In our setting, medical coding requires a high level of accuracy, as the Centers for Medicare & Medicaid Services (CMS) have strict rules on false positives, which can result in severe financial penalties. At the same time, clients are sensitive to false negatives, which lead to loss of reimbursement. Senior workers tend to be more concerned with the potential damage caused by AI.

This reasoning suggests that compared with junior workers, the senior workers would exhibit a tendency to scrutinize AI closely and rely more on their own decisions rather than those of the AI for task completion. Therefore, we expect that human workers with more seniority experience are more likely to resist AI output and benefit less from it than workers with less seniority experience.

3. Research Context

3.1. Background and Task

Healthcare is one of the leading domains for AI applications today (Jha and Topol 2016). Fueled by the adoption of new technologies and widespread digitization of health-related data, the landscape of healthcare has changed rapidly in the past two decades. All major players in the healthcare ecosystem, including government agencies (Talley et al. 2011), healthcare providers (Krittana-wong et al. 2017), insurance companies (Kose et al. 2015), and pharmaceutical manufacturers (Ekins 2016), express enthusiasm about the potential benefits of AI. Given the size (about one fifth of GDP) and nature (extensive knowledge work) of the healthcare industry, the potential of AI in this setting is substantial.

Our research context is the medical coding industry. In the U.S. healthcare system (and in many other countries as well), patient conditions and treatments need to be transformed into standardized codes in the billing process. Accurate medical coding is necessary both for timely and correct payment and for efficient clinical decision making. Historically, medical coding is a labor-intensive job that involves manual code evaluation. It is of considerable economic significance; the market size of the medical coding industry was \$10.6 billion worldwide in 2016 and is increasing 10% per year (Grand View Research 2018).

In our study, we focus on one of the most complex tasks in medical coding—risk adjustment coding from medical charts, which requires human workers to review the complete chart, especially the unstructured physician notes, and make judgments about whether the patient has certain medical conditions, such as

diabetes or cardiac disease. The health conditions identified are designated as *risks* and used to adjust reimbursements (i.e., for the same clinical procedure, reimbursement for treatments received by patients with higher risks will be higher). The industry has widely adopted the Hierarchical Condition Categories (HCC) coding system created by the CMS (Li et al. 2010). The economic value of the coding activity is substantial; an average HCC code has a reimbursement value of several thousand dollars (Pope et al. 2004). Our collaborator is a leading public healthcare analytics company that provides medical chart coding services to multiple insurance companies, with hundreds of coders who have collectively coded over 36 million medical charts in the past decade.

HCC coding is a complex task. Unlike other well-known traditional medical coding tasks, such as International Classification of Diseases (ICD) coding where there is clearly dedicated documentation for diagnoses, the input in HCC coding is the entire medical chart, which is not designed for HCC coding. Moreover, coders do not know *ex ante* if there is an HCC code extractable from the text in the patient chart. Therefore, the decision space is much larger for HCC coders than for ICD coders. Finally, HCC coding needs to exclude known diagnoses and focus rather on unknown chronic conditions; this is an important difference from traditional medical coding that makes HCC coding a complicated and cognitively nonroutine task.

The coding activity proceeds as follows; every time a coder requests a new chart to code, a medical chart is randomly assigned. The medical chart is displayed on the coder's desktop screen and is now ready to be coded. There are three steps for the coder to complete the task. In the first step, the coder will browse the patient chart quickly to form an "image" or a rough understanding of the patient. This will provide a basis for the likelihood of certain conditions (for example, patients who are over 60 years old and overweight tend to have a higher chance of diabetes). In the second step, the coder scans each page of the chart and identifies possible sentences that reflect a condition, similar to a keyword search (for instance, coders may scan for "diabetes" or "HbA1C"). In the third step, complex judgment is needed to decide whether the description confirms the existence of an HCC (e.g., whether a mention of diabetes refers to the specific patient or to family history). Sometimes, the information is subtle. For example, if the patient receives regular insulin injections, the patient has a high likelihood of having diabetes, even though the doctor did not explicitly note the diagnosis.

Based on discussion with experts and the company's management team, we identified the time spent reviewing a medical chart as the measure of a coder's productivity. Although one might think that the HCC codes are the ultimate output, the number of HCC codes that can be detected is not purely driven by coder efforts but is

also determined by the nature of information in the chart. In addition, to ensure quality, for every possible code identified, a coder expends almost as much effort to determine if the code is a false positive. Each coder is subject to postreviews of randomly selected coded charts to minimize coding errors and to ensure quality. According to the company's policy, all coders must maintain over 90% accuracy in their reported HCC findings. Otherwise, the coder will be asked to complete a training program (usually lasting several days) and will not be assigned work for the duration of training. Given that the charts are randomly assigned and the coding quality is well maintained across coders, the average time taken to code a chart reflects a natural measure of productivity.

We note that the practice of medical chart coding is representative of the activities that typical knowledge workers do in other industries. The job is a nonroutine task because HCC codes are not directly included in the medical charts, so coders need to read, understand, and interpret the information in order to decide which HCC codes should be reported. Moreover, every medical chart includes large amounts of patient-specific information, requiring a coder to exercise comprehensive reasoning, judgment, and decision making for every medical chart (Dimick 2010). In contrast, routine coding tasks (such as ICD coding) can be easily performed using technologies like natural language processing. Unsurprisingly, computer-assisted coding technology for ICD coding has become a "must have" in the industry, potentially eliminating human coders' jobs (Crawford 2013).

3.2. Development of the AI

We developed a machine learning-based AI to facilitate the labor-intensive process of risk-adjusted medical chart coding. Specifically, the task that this AI accomplishes is to highlight sentences with potential HCC codes. To do this, the AI first processes all sentences in the chart through a filter. This filter relies on a dictionary developed and maintained by experts in the company to capture all possible keywords that could indicate HCC-related health conditions. However, keyword matching yields too many false positives. In the second step, a machine learning model is deployed to evaluate the probability that the focal sentence contains valid HCC codes and then, to highlight that sentence for the coder to review. This process is illustrated in Figure 1. Among the three steps in HCC coding described in Section 3.1, our AI mainly facilitates step 2 in that it automatically highlights text that has a high chance of containing HCC codes.

For model development, we used 26,000 labeled medical charts, of which 24,000 were randomly selected as training data. After training, the model was tested on the remaining 2,000 charts. We developed several versions of the AI using different approaches, including support vector machine (SVM), convolutional neural networks, and recurrent neural networks. Our implementation is

Figure 1. (Color online) Example of AI Findings in a Medical Chart

(a)

James Doe
Male DOB: 3/21/1932

Previous Tobacco Use: Signed On - 01/01/2014
Smoked Tobacco Use: former smoker
Pack-years: 0
Year started: 1990
Years Since Last Quit: 35 years, 8 months, 9 days
Smokeless Tobacco Use: 0
Counseled to quit/outdown: yes
Passive smoke exposure: no
He does not drink alcohol.

Additional Social History (reviewed - no changes required):
Children: 8 children
Lives with spouse/partner
Retired from being a buyer
Works part time at quit smoking in 1987

Allergies:
* FLOMAX (Critical)
BETA BLOCKERS (PROPRANOLOL HCL) (Critical)

Family History Summary:
Mother (biol) - Has Family History Coronary Heart Disease female < 65 - Entered On: 01/01/2015

General Comments - FH:
Female relative developed heart disease before the age of 65. No male relative developed heart disease before the age of 65.
Mother is deceased, died of cancer at 50. Father is deceased, died of non-cardiac cause at 80.

Social History:
Reviewed history from 08/08/2013 and no changes required:
Children: 8 children
Lives with spouse/partner
Retired from being a buyer
Works part time at LifeWay Christian bookstore
quit smoking in 1987

Review of Systems
General: Complains of fatigue
Cardiovascular: Complains of lightheadedness/dizzy, chest pain or discomfort, shortness of breath with exertion, swelling of hands or feet, difficulty breathing while lying down
Respiratory: Patient denies sputum, wheezing, shortness of breath, excessive snoring, chronic cough
Musculoskeletal: Complains of back pain, arthritis
The remainder of the complete review of systems was negative.

James Doe

**Myocardial Perfusion Imaging
Regadenoson (Lexiscan)**

Patient:	James Doe	DOB:	3/21/1932	Age:	75
MRN:		Height:	172.7 cm	Gender:	M
Account #:	02/01/2015	Weight:	83.2 kg	BSA:	2.14 m ²
Study Date:		Room:	CTCU		

READING: John Smith MD
ADMITTING: Jane Doe MD
ATTENDING: Lisa Davis MD
ORDERING: Dr. Jenn
NUCLEAR TECH: Dr. Wonder
NUCLEAR TECH: David Goodhealth
OTHER: REFERRING

Indication: Jw pain SOB, primary Dr Danner
History: Risk factors, COPD Family history of coronary artery disease, Former tobacco use, Hypertension, Dyslipidemia

Study date: No prior study is available for comparison. Study status: Routine. Objective: Diagnostic evaluation.
Consent: The risks, benefits, and alternatives to the procedure were explained to the patient and informed consent was obtained. Procedure: Initial setup. A baseline ECG was recorded. Intravenous access was obtained. Surface ECG leads and manual cuff blood pressure measurements were monitored. Regadenoson (Lexiscan) stress test. Stress testing was performed, with regadenoson (Lexiscan) by intravenous bolus, for a total dose of 0.4mg over 10.00 sec, followed by a 5 ml saline flush. Exercise testing was performed. Exercise was terminated due to protocol completion. Study completion: All catheters inserted during the procedure were removed. The patient tolerated the procedure well and was discharged from the lab.

Lab, prior tests, procedures, and surgery:
PCI
Myocardial perfusion imaging: Regadenoson (Lexiscan). Gated SPECT and planar imaging. Birthdate: 3/21/1932
Age: Patient is 1. Sex: Gender: male. Ethnicity: white. Height: Height: 172.7 cm. Weight: 88 in. Weight: 183.2 lb. Weight: 205 lb. Body mass index: BMI: 31.2 kg/m². Body surface area: BSA: 2.14 m². Patient status: Outpatient. Study date: Study date: 03/01/2015. Location: Stress laboratory

Baseline ECG: Sinus bradycardia.

Stage	HR	BP (mmHg)	Rhythm	Symptoms	Comments
Baseline	57	100/55 (119)	Sinus brady	None	—
Regadenoson (Lexiscan): 1 min	85	160/85 (117)	NSR, ventricular bigeminy	Headache, mild chest tightness	occasional cough
Recovery: 1 min	81	177/86 (116)	NSR	—	ventricular bigeminy
Recovery: 2 min	77	154/78	—	Subsiding, no cto chest	—

Page 1 / 2

(b)

Review of Systems
General: Complains of fatigue.
Cardiovascular: Complains of lightheadedness/dizzy, chest pain or discomfort, shortness of breath with exertion, swelling of hands or feet, difficulty breathing while lying down.
Respiratory: Patient denies sputum, wheezing, shortness of breath, excessive snoring, chronic cough.
Musculoskeletal: Complains of back pain, arthritis.
The remainder of the complete review of systems was negative.

History: Risk factors: COPD Family history of coronary artery disease. Former tobacco use. Hypertension.

Notes. The highlighting in panel (a) is done by AI, and it is magnified in panel (b). (a) Two example pages. (b) AI findings from two example pages.

based on the SVM version because of its superior computing efficiency.⁴ Our SVM model outputs a probability, which allows us to customize the threshold to control the recall (the portion of the HCC codes captured by the AI). The company set the threshold as 0.90 (recall as roughly 95%) for the best outcome.

Given its superior performance, this AI has substantially increased the level of machine intelligence in the industry. Before the introduction of ML-based AI, both academia and industry had expended significant effort in entity recognition from clinical notes using rule-based models. One of the most famous tools is clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al. 2010). During our model development, we benchmarked our AI models with the performance of cTAKES on our data. Given a level of recall (approximately 90%) similar to that required of our SVM model, the precision of cTAKES is only 6.5%. With such a high false-positive rate, it is not feasible for coders to use rule-based models in real coding work. Our AI achieves a precision of about 30% while maintaining a recall of roughly

95%. With its exclusive ability to handle the cognitively complex task of HCC coding, it is making a substantial difference in coder performance. Putting the performance in context, the improvement of precision from 6.5% to 30% is significant. A precision of 6.5% means coders must verify 15 findings (detected by the system) to isolate one HCC code. When coders use our machine learning system, this number is reduced to about three findings. This is a significant distinction in practice, and the company finds the resulting improvement striking. The company has implemented it in daily practice, and executives attribute significant revenue generation to its use.

We note that this AI system is representative of current state-of-the-art applications of AI in knowledge work; the AI assists the human who makes the final decisions. For example, one of the most successful use cases of AI in medicine is diagnoses from imaging, where AI suggests the diagnosis and human doctors make the final determination (Hainc et al. 2017, Hosny et al. 2018). Similarly, the AI developed for this study highlights sentences where a code might be identified and suggests the

HCC code that may apply, but it still requires human coders to review its findings. As noted by industry experts (Williams 2015), the singularity of AI is still decades away, and the majority of current AI applications are similar to our use case, in which AI's reasoning facilitates rather than completely replaces human decision making. Therefore, findings from this study are pertinent to the current practice of AI use in business.

One point worth noting is that our study setting is less confounded with resistance to learning new technologies, which is a common confounding factor with human capital. First, there was no substantive change in the workflow. Before implementation of our AI, the coders in our study context were all using a computer interface to review medical charts and report coding results. The AI only provides highlights on coders' screens when they review the medical chart. The entire system stayed the same for coders, and all of the AI's reasoning occurs in the background without any interference with coders' workflow. Given that coders were already fluent with the computer system, there may be no cause for concern regarding adoption.

The AI system at the center of our study went online in July 2018. With help from management, 80 coders, selected to represent the full spectrum of coding seniority levels,⁵ received the AI. The remaining 468 coders who did not have access to the AI constitute our control group. To help understand the impact on productivity, we collected data for a year before the AI was used. This pretreatment period runs from July 16, 2017 to April 30, 2018 (May and June 2018 were excluded because of the adjustment for transition). We also collected coder performance data from July 1, 2018 to October 31, 2018, which is defined as the posttreatment period.⁶ During our study period, no major changes were made to the work procedures aside from the implementation of AI. The 80 coders in the treatment group reviewed and coded 195,732 charts. The control group (468 coders) coded 925,099 medical charts in the same period for a total of 1,120,831 medical charts in the study sample.⁷

3.3. AI's Impact on HCC Coding

We now contextualize our theoretical arguments for the interplay between AI and worker experience in the specific task of HCC coding (task experience). Because the aim of HCC coding is to find certain chronic conditions that were not reported in the structured data, only a very small number of HCC codes can be found in each medical chart, and a large portion of medical charts contain zero HCC codes. Therefore, human coders expend the most time and effort confirming the negative labels of certain sentences. Over years of practice, coders gain experience and learn how to separate the wheat from the chaff. AI has many strengths that could facilitate coders' work, and there are reasons to believe that AI could potentially substitute for a coder's valuable experience.

First, the AI was trained on 24,000 charts and learned how sentences containing HCC conditions are systematically different from irrelevant ones. It can further *conservatively* rule out negative cases. With AI's help, coders could spend less time on irrelevant information collection and conduct fewer examinations of negative cases, which is especially beneficial to workers with less experience. Second, human coders tend to skip sections they deem irrelevant, and this may lead to important clues being overlooked. The AI is indefatigable and can tirelessly check each sentence. Therefore, the highlighting provided by the AI helps prevent reviewers from making faulty determinations of irrelevance because of negligence. This is naturally more helpful to workers with less experience because of their higher tendency to recklessly skip uncommon but suspicious sections. Third, the AI's knowledge comes from the training data, which it utilizes objectively to make decisions, in contrast to human coders who may be prone to many subjective biases, such as anchoring (Tversky and Kahneman 1974) and recency effects (Tzeng 1973). Although more experienced coders who learned from feedback are less prone to such biases, AI again assists less experienced coders more, with its objectivity correcting for coders' biases. All these mechanisms make experience less valuable in the presence of AI.

On the other hand, experience can complement AI in medical coding, as theorized in Section 2.4.1. In the initial step of the process, human coders can quickly read through the medical chart to develop familiarity with and an overall understanding of the patient's situation (e.g., "the patient is overweight and lacks exercise and therefore has a high risk of being diabetic"). Based on this comprehensive overview of the patient, a human coder then zooms in to search for evidence of suspicious descriptions of possible chronic conditions. Here, AI is at a serious disadvantage as it lacks the tacit knowledge or "common sense" to form a comprehensive understanding of the patient. This is, again, typical of the current state-of-the-art AI in knowledge work. The AI simply mechanically analyzes each sentence and computes the probability that an HCC is included. Therefore, the AI tends to produce more false positives than a human. Given this, although all coders will benefit from the help of AI, it is likely that coders with rich task experience can critically evaluate AI's output and rule out the false positives more effectively than less experienced coders. The sharper experience related to how and when to use commonsense knowledge (and the stronger commonsense medical knowledge accumulated through task experience) can further help more experienced coders outperform their peers in leveraging AI. This reasoning is consistent with the trend of IT as an SBTC in the past (Krusell et al. 2000, Bresnahan et al. 2002). Therefore, although both substitution and complementarities might theoretically exist, we expect the latter to dominate given

that HCC coding is a nonroutine and cognitively complex task.

The role of seniority (time-based) experience in human-AI teaming and its interaction with human workers are more complicated, as other organizational and psychological factors also come into play. Although senior workers are possibly better able to complement AI (following the logic of SBTC), a countervailing force is that more senior workers may have less trust in AI because they doubt its competence. First, because of AI's limited common sense and less than perfect performance, human coders, similar to a supervisor, have to intervene and correct the errors made by the AI. Senior coders are more likely to detect and observe the shortcomings of AI, which can erode their trust in the technology. Second, the black-box characteristic of AI remains a significant barrier to human-AI collaboration. In the HCC coding context, the false-positive highlights from AI likely create more confusion for more experienced coders who because of the AI's opaque reasoning process, find it harder to understand its underlying logic, further raising concerns about integrity and reducing trust. Finally, senior coders' careers are more closely tied to the overall performance of the medical coding activity. Higher sensitivity to the shortcomings of AI coupled with the opaqueness behind its reasoning likely increase their concerns about AI's potential damage to the organization, further dampening trust.

The trust deficit among senior coders means that they are more likely to doubt AI, spend more time checking the AI's recommendations, discard HCC codes found by the AI, or to be extra cautious, check the portions of the charts that are marked as negative by AI. In doing so, senior workers may forgo the advantages offered by AI: reading less, coding less, and reallocating their regained time to focus on sentences where human intervention and judgment are required.

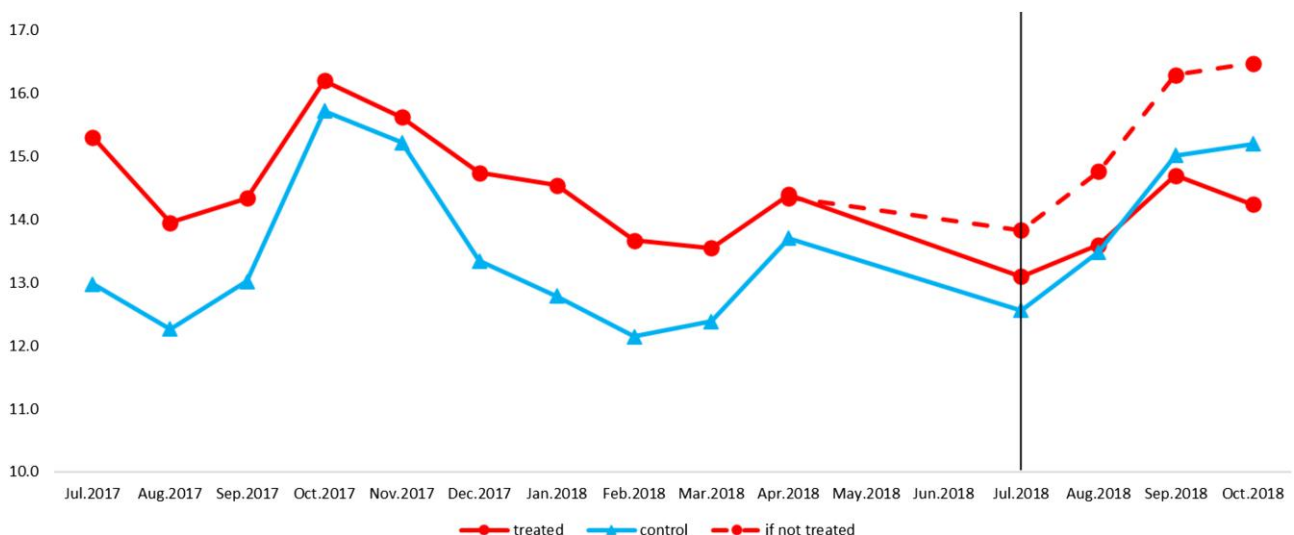
4. Results: Impact of AI on Productivity

4.1. Overall Impact of AI on Productivity

We use all the medical charts coded by coders in the two study groups to show the model free evidence. There are 1,120,831 medical charts in the full study sample, with 195,732 medical charts in the treatment group and 925,099 medical charts in the control group. Pooling both groups together, the average time to code one medical chart is 13.62 minutes with a standard deviation of 28.35; the average number of pages is 35.31 with a standard deviation of 50.12. Focusing only on treated medical charts, the average time spent on one medical chart is 14.37 minutes with a standard deviation of 27.01; the average number of pages is 34.81 with a standard deviation of 50.91.

Figure 2 plots the average time (in minutes) that it took coders in both the treatment and control groups to code one medical chart. As shown in the figure, the control group and treatment group show similar fluctuation in trends in the preperiod (July 2017 to April 2018). During this period, coders in the treatment group spent 1.27 more minutes on each medical chart than coders in the control group.⁸ Leveraging this average difference between the two groups in the pretreatment period, we can calculate the reduction of coding time because of AI implementation. The dotted line in Figure 2 shows the time that the treatment group would have spent on an average medical chart if that 1.27-minute difference had remained consistent throughout the study. Comparing this dotted-line projection with the observed trend, we see that AI reduced the coding time for an average medical chart by 1.16 minutes (7.84%) in August, 1.59 minutes (9.77%) in September, and 2.24 minutes (13.59%) in October. We conclude that across the three months in the postperiod, coding time was reduced by 1.28 minutes (9.17%) per medical chart after AI implementation.

Figure 2. (Color online) Trends in Medical Chart Coding Time



The treatment group and the control group follow a parallel trend but do not exactly overlap in the pre-AI period (Figure 2). Although the parallel trend is a valid basis for inferring causality (Card and Krueger 1993), concerns may be raised about the difference between the treatment and control coders. Therefore, we further conduct exact matching at the medical chart level by selecting medical charts that have similar values on all control variables, including number of pages per chart, time of day of chart coding, code type, and coders' seniority. Collectively, these variables are related to both the volume of work and work patterns, which helps ensure equivalence by eliminating the variance attributable to them. Our matching procedure balances the number of medical charts completed by the two groups (treatment and control) and in the two study periods (pre and post). To ensure the matched group represents equivalent coders, we also remove coders who have started the coding work in the postperiod (zero medical charts completed in the preperiod) or who stopped coding before the postperiod (zero medical charts completed in the postperiod). The matched sample has 64,280 charts with a balanced number of charts across the pre- and postperiods, as well as control and treatment groups.

The balance check and summary statistics of the matched data are reported in Table 1. After matching, key variables, including number of pages, coding time of the day, type of coding, and the two types of experience, show no statistically significant differences in both the pre- and postperiods. As a robustness check, we also report analyses using the full sample in the online appendix. All the results are consistent.

Equation (1) depicts our formal empirical specification. To eliminate potential confounders, we conduct

coder-level fixed effects analysis. In our model, the dependent variable (Y_i) is the time (in minutes) spent on a medical chart. *Post* is a dummy variable that takes the value of zero for the pre-AI period and one for the post-AI period. *AI* (omitted in the fixed effects model) is one for the treatment group (i.e., those 80 coders who were assigned to use the AI) and zero for the remaining coders who did not use AI. The interaction term $Post \times AI$ captures the effect of AI on review time:

$$Y_i = \beta_0 + \beta_1 Post \times AI + \beta_2 X_i + FE_{corder} + FE_{month} + \varepsilon_i. \quad (1)$$

We also include a set of medical chart-level characteristics as control variables. We control for the time of day coding is performed by including three dummy variables: morning, afternoon, and night. The length of medical charts is controlled for using the number of pages (*NumPage*). Finally, the type of coding (i.e., different versions of HCC codes from the CMS and the U.S. Department of Health and Human Services (HHS)), the round of coding (some charts are randomly chosen by the company for a second round of coding for quality control), and month when coding was performed are also incorporated into our model. Individual coder characteristics are controlled for using fixed effects.

The estimation of AI's impact on productivity is reported in Table 2. We use an ordinary least squares (OLS) model with chart-level controls and coder fixed effects with chart-level controls (column (1)). As reported in column (1), the coefficient of $Post \times AI$ is -1.17 , which is statistically significant ($p < 0.01$). On average, AI reduces coding time by 1.17 minutes (11.29% in the pre-treatment period) per medical chart.

Concerns might be raised about whether the productivity increase was because of a reviewer's rush to

Table 1. Summary Statistics and Balance Check of the Matched Data

	Pre			Post		
	Control	Treated	Difference	Control	Treated	Difference
<i>Number of Charts</i>	16,070	16,070	—	16,070	16,070	—
<i>Review Time (in minutes)</i>	10.46 (10.43)	10.36 (11.04)	$p > 0.1$	11.23 (12.73)	7.48 (6.72)	$p < 0.01$
<i>Number of Pages</i>	30.70 (38.46)	31.20 (48.73)	$p > 0.1$	25.17 (24.99)	24.78 (30.32)	$p > 0.1$
<i>Percentage of Charts Finished in Morning</i>	17.48% (37.99%)	17.49% (37.99%)	$p > 0.1$	24.93% (43.26%)	24.74% (43.15%)	$p > 0.1$
<i>Percentage of Charts Finished in Afternoon</i>	39.88% (48.97%)	39.85% (48.96%)	$p > 0.1$	39.99% (48.99%)	40.67% (49.12%)	$p > 0.1$
<i>Percentage of Charts Finished at Night</i>	42.64% (49.46%)	42.66% (49.46%)	$p > 0.1$	35.08% (47.72%)	34.59% (47.57%)	$p > 0.1$
<i>Percentage of CMS HCC Coding (vs. HHS HCC Coding)</i>	53.39% (49.89%)	53.13% (49.90%)	$p > 0.1$	73.21% (44.29%)	73.80% (43.98%)	$p > 0.1$
<i>Percentage of Senior Coders</i>	29.25% (45.49%)	28.78% (45.28%)	$p > 0.1$	27.02% (44.41%)	27.64% (44.72%)	$p > 0.1$
<i>Seniority Exp</i>	60.85% (48.81%)	60.09% (48.97%)	$p > 0.1$	46.68% (49.89%)	45.99% (49.84%)	$p > 0.1$
<i>Percentage of Senior Coders Task Exp</i>	13,761.45 (4,664.63)	13,838.03 (5,216.99)	$p > 0.1$	12,255.74 (4,753.54)	12,325.75 (5,428.99)	$p > 0.1$
<i>Number of Charts Ever Reviewed</i>	2,011.72 (3.37)	2,011.67 (3.14)	$p > 0.1$	2,011.47 (3.17)	2,011.43 (3.06)	$p > 0.1$
<i>Start Year</i>						

Note. Standard deviations are in parentheses.

Table 2. AI Impact on Medical Chart Coding Time

	(1) Main result (OLS)	(2) Main result (OLS)	(3) Main result (OLS)	(4) Main result (Poisson)
Dependent variable	Review Time	Review Time	NumHCC	NumHCC
$Post \times AI$	−1.17*** (0.23)	−1.14*** (0.24)	−0.01 (0.02)	−0.11 (0.07)
NumHCC		4.39*** (0.15)		
Constant	7.31** (3.08)	2.65 (1.95)	1.06*** (0.37)	
Control variables	NumPage, NumHCC, Round of Coding, Type of Coding, Time of Day			
Observations	64,280	64,280	64,280	63,930
Coder fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
R^2	0.23	0.28	0.05	
Number of coders	361	361	361	340

Notes. Poisson regression model is used for column (4). Robust and clustered standard errors are in parentheses.

** $p < 0.05$; *** $p < 0.01$.

complete the job, thereby lowering output in terms of the number of codes extracted. We, therefore, further control for the number of HCC codes found in the medical chart and estimate the model again. The coefficient of $Post \times AI$ is again negative and significant (−1.14, see column (2)), further affirming the result.

We demonstrate a significant improvement in medical chart coding productivity because of the use of AI. As previously mentioned, one might be concerned that the increased productivity comes at the cost of a decrease in output; that is, coders might speed up their work but identify fewer HCC codes. To uncover the impact of the AI on output quality, we modify Equation (1) by using the number of detected HCC codes as the dependent variable. This result is reported in column (3) of Table 2. $Post \times AI$ has an insignificant coefficient, which means that AI does not lead to a deterioration in quality. We estimate a Poisson specification as a robustness check because the dependent variable is the count of found HCC codes. As reported in column (4), the result is consistent. Moreover, coding practice in the field also confirmed a stable quality of the reported HCC codes. First, several institutional factors ensure the stability of the service quality in terms of type 1 and type 2 errors. Given the high economic value of each HCC code, the clients of our collaborator often give the same batch of charts to its competitors and compare the outcome. If the AI misses valid HCC codes (type 2 error), that will cause concerns from the clients. This has forced the company to have a robust process to ensure the consistency of quality. The CMS will further evaluate the reported HCC codes to decide whether they are valid. If yes, reimbursement will be made to the providers by the CMS. If not, they will not be reimbursed, and the entire claim will be returned. The CMS is especially sensitive to up-coding cases where

HCC codes were falsely identified (type 1 error), which could lead to the disqualification and even penalty. Given this robust evidence, we can rule out the possibility that the AI affects the quality of work by a meaningful magnitude.

In summary, we find evidence of a positive overall effect of AI on productivity. Next, we examine how these effects are heterogeneously distributed across different types of experience.

4.2. Task Experience and the Impact of AI

We focus first on task experience, which as argued theoretically and confirmed in our data, is not significantly positively correlated with seniority experience (the correlation coefficient is 0.11). In other words, senior coders may have been in their positions longer but may also have completed fewer medical charts in total. This is mainly because of the increased possibility of being assigned to other responsibilities as their organizational tenure grows. The divergence provides a perfect context for the separation of two types of experience. As described in Section 2, we define coders' task experience as the number of medical charts they have completed in their entire job history. We use the average number of medical charts completed across all coders as the threshold (which is 12,020) to label high and low task experience coders. In the treatment group, 27 coders are classified as high task experience, whereas the remaining 48 coders are classified as low task experience. In the same period, for the control group, 135 coders are classified as high task experience, whereas the remaining 151 coders are classified as low task experience. We create a dummy variable for each task experience level and then interact the two level dummies with the treatment effect $Post \times AI$. To formally test the interaction between the AI

and workers' task experience, we use the model specified in Equation (2):

$$Y_i = \beta_0 + \beta_1 Post \times HighTaskExp + \beta_2 Post \times LowTaskExp + \beta_3 Post \times AI \times HighTaskExp + \beta_4 Post \times AI \times LowTaskExp + \beta_5 X_i + FE_{coder} + FE_{month} + \varepsilon_i. \quad (2)$$

The results are reported in Table 3. We find that the moderating effect of high task experience is negative and statistically significant ($p < 0.01$, column (1)), implying that AI helps coders with high task experience shorten chart review time and improve productivity. Although coders with high task experience can benefit significantly more from AI, the coefficient of low task experience's moderating effect is positive, although insignificant, suggestive of a complementarity. This is further confirmed in subgroup analyses, where the treatment effect is significant for high task experience coders ($p < 0.01$, column (2)) but not for low task experience coders ($p > 0.1$, column (3)). This finding supports the conjecture of complementarity between AI and task experience.

In the analyses, we used the average number of medical charts reviewed by all coders as the threshold to define high and low task experience. To validate the robustness of our finding, we use the continuous measure (mean centered) of task experience and then conduct the same moderating analyses. As reported in

column (4), $Post \times AI \times NumChartEver$ is consistently significant ($p < 0.01$).

In any form of collaboration, the team typically performs better than individual members because of the different and unique contributions made by each member. In a team consisting of members who all make zero contribution or each makes the full contribution, there is no complementarity. In the intermediate case, the more unique contributions each member makes, the more complementarities are created. For a cognitively complex task like HCC coding, neither human coders nor contemporary AI are either useless or perfect. In this case, two observations about our finding of the complementarity between AI and task experience are noteworthy. First, we confirm the different and unique advantages possessed by both human workers and AI in this teaming. The finding is consistent with recent human-AI teaming literature, where AI recognized to be advantageous at computational skills and humans makes unique contributions based on their broad knowledge (experience) (Fügener et al. 2021, Luo et al. 2021). Second, a positive coefficient of high task experience supports the value of human experience in teaming with AI. In other words, the accumulation of task experience brings in unique contributions that cannot be covered by AI. This is logically plausible given that the AI literature clearly specifies different learning mechanisms in AI from human learning (LeCun et al. 2015). As a result, more

Table 3. AI Benefit to Workers at Different Task Experience Levels (with Coder Fixed Effects)

	(1) <i>Moderator TaskExp</i>	(2) <i>Within highTaskExp</i>	(3) <i>Within lowTaskExp</i>	(4) <i>Continuous TaskExp</i>
Dependent variable	Review Time	Review Time	Review Time	Review Time
$Post \times AI$		−1.71*** (0.32)	−0.26 (0.27)	−1.08*** (0.22)
$Post \times highTaskExp$	0.08 (0.23)			
$Post \times AI \times highTaskExp$	−1.72*** (0.32)			
$Post \times AI \times lowTaskExp$	−0.31 (0.27)			
$Post \times AI \times NumChartEver$				−0.11*** (0.04)
$Post \times NumChartEver$				0.02 (0.03)
Constant	2.59 (1.99)	3.14* (1.61)	2.56 (2.80)	2.65 (1.97)
Control variables	<i>NumPage, NumHCC, Round of Coding, Type of Coding, Time of Day</i>			
Observations	64,280	34,327	29,953	64,280
Coder fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
R^2	0.28	0.27	0.30	0.28
Number of coders	361	162	199	361

Notes. We used per 1,000 chart for *NumChartEver*. Robust and clustered standard errors are in parentheses.

* $p < 0.1$; *** $p < 0.01$.

task experience leads to more human learning that is beyond AI's learning ability, which could further complement AI in task execution.

4.3. Seniority Experience and the Impact of AI

If AI and experience are complementary, one would expect senior workers to benefit more from AI than their less experienced colleagues, a conjecture supported by the task experience analysis. However, alternatively, as discussed in Sections 2.3 and 3.3, organizational and psychological factors might prevent senior coders from teaming with AI. We now report analyses to test these competing conjectures. We measure workers' seniority experience based on their tenure (number of years in the coding job) and classify coders into senior (10 or more years of experience) and junior (less than 10 years of experience) categories.⁹ In the treatment group, there are 26 senior coders and 49 junior coders. Correspondingly, in the control group, there are 47 senior coders and 239 junior coders. We create a dummy variable for each seniority level and then interact the two seniority-level dummies with the treatment effect $Post \times AI$.

We replace high task experience and low task experience in Equation (2) with senior and junior measures to

conduct the moderating analyses:

$$Y_i = \beta_0 + \beta_1 Post \times Senior + \beta_2 Post \times Junior + \beta_3 Post \times AI \times Senior + \beta_4 Post \times AI \times Junior + \beta_5 X_i + FE_{coder} + FE_{month} + \varepsilon_i. \quad (3)$$

Results are reported in Table 4. As shown in column (1), we find the moderating effect for junior coders to be negative and significant ($p < 0.01$), which suggests that AI helps junior workers to shorten chart review time and improve productivity. The moderating effect of senior coders is also negative but smaller in magnitude, and it is not statistically significant. This finding suggests that AI helps junior coders more in improving productivity. We also confirm this result in the subgroup analysis for senior coders and junior coders. As shown in columns (2) and (3), the treatment effect is significant for junior coders ($p < 0.01$, column (3)) but not for senior coders ($p > 0.1$, column (2)).

Although the company's managers suggested 10 years of experience as the threshold for senior coders, we conduct further analysis to reinforce the generalizability of the findings about seniority. We use a continuous measure in number of years (mean centered), and we estimate the same regression model as before to examine how seniority level influences the AI's effects. The results

Table 4. AI Benefit to Workers at Different Seniority Levels (with Coder Fixed Effects)

	(1) <i>Moderator Senior</i>	(2) <i>Within Senior</i>	(3) <i>Within Junior</i>	(4) <i>Continuous Seniority</i>	(5) <i>Continuous Two Types</i>
Dependent variable	Review Time	Review Time	Review Time	Review Time	Review Time
$Post \times AI$		−0.43 (0.45)	−1.45*** (0.27)	−1.16*** (0.24)	−1.08*** (0.22)
$Post \times Senior$	−0.04 (0.26)				
$Post \times AI \times Senior$	−0.39 (0.43)				
$Post \times AI \times Junior$	−1.44*** (0.27)				
$Post \times AI \times NumYears$				0.17** (0.08)	0.23*** (0.08)
$Post \times NumYears$				−0.04 (0.03)	−0.03 (0.03)
$Post \times AI \times NumChartEver$					−0.14*** (0.04)
$Post \times NumChartEver$					0.02 (0.03)
Constant	2.60 (1.94)	1.65 (2.59)	3.40 (2.80)	2.64 (1.92)	2.63 (1.94)
Control variables	<i>NumPage, NumHCC, Round of Coding, Type of Coding, Time of Day</i>				
Observations	64,280	18,109	46,171	64,280	64,280
Coder fixed effects	Yes	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes	Yes
R^2	0.28	0.37	0.26	0.28	0.28
Number of coders	361	73	288	361	361

Notes. We used per 1,000 chart for *NumChartEver*. Robust and clustered are standard errors in parentheses.

** $p < 0.05$; *** $p < 0.01$.

are reported in column (4) of Table 4. The coefficient of $Post \times AI \times NumYears$ is significant ($p < 0.05$).

It is natural to assume that high seniority experience leads to higher task experience. However, the correlation coefficient between task volume and organizational tenure is only 0.11 among all coders in this study. The divergence of the two experience measures could be explained by organizational factors as summarized by Tesluk and Jacobs (1998) and Ng and Feldman (2010). Although some coders focus on working on the same task over the years, other coders might be assigned diverse tasks, such as training juniors, auditing quality, management, and meeting with clients. To statistically examine if our findings are affected by any potential correlations between the two experience measures, we further combined the two experience measures in one model to test their moderating effects simultaneously. As reported in column (5) of Table 4, task experience complements AI significantly positive in increasing the productivity. Meanwhile, seniority experience remains significant and negative in enhancing the benefits of AI.

5. Mechanisms Underlying Differences in Teaming with AI

5.1. Resistance to AI

Our finding that AI is more helpful for high task experience workers supports the theoretical conjecture that AI complements, rather than substitutes for, human worker experience. However, if AI and human capital are truly complementary, it is puzzling why senior workers with diverse responsibilities tend to benefit less from AI. In this section, we report additional qualitative analyses conducted to further uncover the mechanism of AI's impact on senior coders.

We first collect qualitative feedback from senior coders through focus groups and a formal survey after the AI was implemented. We learned that senior coders do not trust AI to perform as well as humans and tend to focus on the errors made by the AI. Senior coders also complained more about errors than junior coders did. Indeed, one senior coder commented that

I don't fully trust the tool to identify codes. I haven't been told if it is supposed to highlight knowns or not so when I see a known not highlighted. I question if the tool is working correctly.

Also, given that the senior coders have greater breadth of experience, they are more likely to detect imperfections in the AI output, which further deteriorates their trust in the AI.

Many areas of the record are highlighted that are not appropriate for coding. Once one area of the record, whether or not appropriately, is highlighted, I need to review the entire record. I have not found this to be helpful.

Recent studies report similar insights: that managers see AI differently (Lebovitz et al. 2021). In explaining the errors made by AI in the field, managers implicate AI's learning mechanism. Managers believe that AI learns from labels to capture know-what knowledge. However, AI cannot achieve the same level of intelligence as experts who have rich know-how knowledge in practice. In our specific context, the company has a wide spectrum of tasks related to medical coding. Senior coders can be assigned to management and training activities to optimize the overall performance of the entire production line. The management, audit, training, and other related work could limit senior coders' time allocation to individual medical chart coding. With the company's long history of practicing in this industry, coders can also be promoted to managers; as coders accumulate organizational tenure, they have more opportunities to be involved in management and build greater loyalty, thereby becoming more sensitive and alert to mistakes in AI output.

To further verify this explanation, we obtained the job titles of coders in our study from the company and classified these coders as manager or nonmanager.¹⁰ We also use manager as a proxy of seniority experience to conduct the same analysis. As reported in Table 5, consistent with our main seniority experience measure, managers show no significant complementarity with AI ($p > 0.1$, columns (1) and (2)), whereas nonmanagers are able to effectively leverage AI and experience a significant boost in productivity ($p < 0.01$, columns (1) and (3)). This evidence further supports the argument that the employee's position in the organization reflects a core element of the seniority measure.

The comments from the qualitative study suggest that because of their low trust in AI, these senior coders opt to review all the information in the charts rather than solely focusing on the areas highlighted by AI. Their resistance is also supported by an additional small-scale laboratory study we conducted in the company's work setting. The nine coders in this laboratory study were typical coders recruited from the coding team, and they had no prior exposure to the AI. They were asked to independently code 100 preselected charts (randomly selected from the medical chart pool). The coders were randomly assigned into three groups (three coders per group) that used different coding methods. Group 1 was the control group without AI, which involved reading through the whole medical chart to find HCC codes, replicating the standard coding practice before AI implementation. The instructions for Group 2 (resistance) were designed to mimic the case of coders resisting AI; although AI findings were provided to them, these coders were told to not rely on the AI results but to still review all the information in the chart. Group 3 was designed to replicate the scenario in which coders have trust and work cooperatively (i.e., coders team with AI).

Table 5. AI Benefit and Manager Status

	(1) Moderator Manager	(2) Manager	(3) nonManager
Dependent variable	Review time	Review time	Review time
<i>Post</i> × <i>Manager</i>	−0.55 (0.35)		
<i>Post</i> × <i>AI</i> × <i>Manager</i>	−0.71 (0.47)		
<i>Post</i> × <i>AI</i> × <i>nonManager</i>	−1.14*** (0.25)		
<i>Post</i> × <i>AI</i>		−0.41 (0.54)	−1.14*** (0.25)
Constant	2.65 (1.95)	5.79*** (0.69)	2.71 (1.95)
Control variables	<i>NumPage, NumHCC, Round of Coding, Type of Coding, Time of Day</i>		
Observations	64,280	2,185	62,095
Coder fixed effects	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes
<i>R</i> ²	0.28	0.53	0.28
Number of coders	361	12	349

Note. Robust and clustered standard errors are in parentheses.

****p* < 0.01.

Coders in Group 3 were instructed to only validate the AI findings, and the validated HCC codes constituted their final coding result.

Results are reported in Table 6. Group 3 coders, whose instructions required them to trust the AI output, achieved much higher productivity ((21.78 – 8.45)/21.78 = 61.2% less coding time) than Group 2 (resistance) coders. Group 3 also demonstrated 45.0% ((15.37 – 8.45)/15.37) more productivity than the control group. This result confirms that AI can significantly improve productivity if coder trust in it is high. It is also noteworthy that the average medical chart coding time of Group 2 is much higher (41.7%) than the control group. This demonstrates one critical point; AI is not necessarily beneficial. In fact, resistance to AI could lead to a negative impact on productivity. The findings from this laboratory study underscore what is widely known in relation to other types of technologies; user acceptance is crucial to leveraging AI for high productivity.

Lastly, we note that in this laboratory study, we sought to create contexts with sharp differences; that is,

coders were instructed to fully cooperate with or fully resist the AI. In reality, coders' resistance to AI more likely exists on a continuum between full cooperation and full resistance. The variation in level of resistance might explain why some senior coders in our main study still benefit from AI but less so than junior coders.

Overall, evidence from the qualitative feedback, our additional regressions, and the laboratory study collectively indicates that resistance to AI plays an important role in workers' inability to realize productivity gains from its use. Because senior coders tend to believe that they have more expertise than the AI and possess greater confidence in their own judgment, they exhibit high resistance. Their concerns about the negatives of AI because of the undue emphasis on its shortcomings and opacity also reduce their ability to effectively team with AI.

6. Falsification Check

To further confirm that our results are driven by AI rather than by common environmental factors that might

Table 6. Verifying the Mechanism of Low Productivity Because of Resistance

	Group 1: Control group			Group 2: No trust in AI			Group 3: High trust in AI		
Coder	1	2	3	4	5	6	7	8	9
<i>Coder tenure</i> (years)	7	4	6	7	2	9	5	7	9
<i>Per chart time</i> (minutes)	15.77	11.43	18.91	18.97	22.96	23.4	9.14	7.98	8.29
<i>NumHCC per chart</i>	1.54	1.62	1.68	1.64	1.61	1.66	1.54	1.46	1.54
<i>Num of charts</i>	100	100	100	100	100	100	99	98	99
<i>Num of AI findings</i>	—	—	—	1,095	1,095	1,095	1,095	1,084	1,095
<i>Per chart time</i> (minutes)		15.37			21.78			8.45	
<i>NumHCC per chart</i>		1.61			1.64			1.51	

affect coding in general (such as new organizational policies, changes in management, etc.), we conduct a falsification test. In addition to HCC coding, the company also works on other chart coding tasks, such as Clinical Risk Group, Healthcare Effectiveness Data and Information Set, and Chronic Illness and Disability Payment System coding. Because our AI was developed to improve HCC coding, these non-HCC coding tasks should not be affected by the AI. Therefore, they serve as good candidates for the falsification.

During our study period, 246,852 of these non-HCC medical charts were coded by all coders (52,208 by the treatment group and 194,644 by the control group). We use the same regression model (Equation 1) for these non-AI coding tasks; results are in Table 7. In column (1), the coefficient of $Post \times AI$ is insignificant, indicating that worker productivity in non-AI coding tasks did not increase in the same period. Also, the moderating effects of experience levels are not significant (columns (2) and (3)), further confirming the insignificant influence of AI implementation on non-AI coding tasks. Therefore, our main finding is likely not because of factors unrelated to the AI.

7. Discussion and Conclusion

AI is one of the most significant technological advances in history, with profound implications for economies

and societies. As a result of the learning capabilities of contemporary AI, there have been substantial concerns about AI's implications for human capital and the labor market. What is the role of human experience in human-AI teaming? In this empirical study in a knowledge-intensive work setting, we build upon and extend the existing literature to examine two types of experience, revealing that they play distinct roles. Highly task-experienced coders enjoy greater improvements in productivity. However, when we turn to a different experience measure using seniority, we surprisingly find that junior coders experience greater gains in teaming with AI than senior coders. We attribute this to senior workers' greater sensitivity to AI imperfections, which results in their resistance to AI use and therefore, lower benefits for them. Interestingly, this effect does not show in greater task experience.

Our findings provide theoretical and empirical foundations for human-AI teaming by generating new insights into the interplay between AI and human experience. First, contrary to the common understanding that AI will make human experience less valuable or even obsolete, our findings regarding task experience support AI's role as a complement to rather than a substitute for human experience. In this regard, our work suggests that AI, at its current stage, remains an instance of SBTC. Second, our study signifies the complexity in human-AI interactions in the teaming relationship, which critically depends on the type of experience; greater task experience

Table 7. Falsification Checks (with Coder Fixed Effects)

	(1) Other Coding Types	(2) Other Coding Types—Seniority Level	(3) Other Coding Types—TaskExp
Dependent variable	Review Time	Review Time	Review Time
$Post \times AI$	2.99 (2.07)		
$Post \times Senior$		−1.84 (1.41)	
$Post \times AI \times Senior$		7.47 (4.90)	
$Post \times AI \times Junior$		1.75 (2.19)	
$Post \times highTaskExp$			2.17 (1.43)
$Post \times AI \times highTaskExp$			4.38 (2.83)
$Post \times AI \times lowTaskExp$			2.00 (2.77)
Constant	8.42 (6.26)	8.67 (6.24)	7.79 (6.28)
Control variables	NumPage, NumCodes, Review Time, Round of Coding, Type of Coding, Time of Day		
Observations	246,852	246,852	246,852
Coder fixed effects	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes
R^2	0.17	0.17	0.17
Number of coders	392	392	392

Note. Robust and clustered standard errors are in parentheses.

does not necessarily lead to algorithm aversion, but greater seniority experience could hamper the effective use of AI. In this way, we advance our understanding of the role of experience in human-AI teaming (Allen and Choudhury 2022) and contribute to this emerging stream of literature (Clement et al. 2020, Guo et al. 2020, Bai et al. 2022). Although Luo et al. (2021) and Allen and Choudhury (2022) mainly focus on the experience of workers, this study advances understanding of the interaction between AI and different types of human experience. We show that experience is multidimensional and that different dimensions have distinct interactions with AI and play unique roles in human-AI teaming. Allen and Choudhury (2022) theoretically argue that in human-AI collaboration, human experience is associated with two driving forces: ability and aversion. We are able to delineate the mechanism by showing that different types of experience have opposite effects on human-AI collaboration. Our task-based experience reflects the concept of ability, and our seniority-based experience is related to the aversion effect. The separation of different types of human experience in this study thus provides a deeper explanation and potentially reconciles previous findings. Additionally, our work contributes to research at the nexus of healthcare and IT. Information systems researchers have studied the impact of IT in healthcare (Ganju et al. 2016, Appari et al. 2018), and the human ability to utilize technology is a core focus of this literature (Atasoy et al. 2017, Karahanna et al. 2019). Our study highlights the need to better understand the potential of AI in the healthcare field.

These findings offer practical implications for the successful use of AI in organizational settings. First, our research provides empirical evidence of the positive effect of human-AI teaming on knowledge worker productivity (Brynjolfsson et al. 2018b, Lou and Wu 2021). Despite the optimism surrounding AI, empirical evidence for its positive effects on economic productivity remains scant for cognitively nonroutine tasks (Case and Deaton 2017, Syverson 2017). This disappointing reality has been dubbed “the AI productivity paradox” (Brynjolfsson et al. 2018b). Unlike recent studies on the impact of algorithms in online platforms and customer decision making (He et al. 2015, Brynjolfsson et al. 2018a, Sun et al. 2021, Bundorf et al. 2022), our setting is a typical knowledge-intensive job with clear measures of productivity. Our finding that AI leads to productivity gain should encourage business leaders and policy makers who are skeptical about its potential to consider investing in AI.

Second, our results may help ease concerns about potential job losses for knowledge workers predicted by forecasters (Cellan-Jones 2014, Davenport and Ronanki 2018). In 2020, the knowledge worker population exceeded 1 billion globally (Ricard 2020). The effect of the coming wave of AI on these jobs thus carries substantial economic and societal significance. Our finding that AI

complements knowledge workers’ task experience may be welcome news to those who are concerned about the negative impact of AI on human capital, which has important policy implications.

Third, from a technological perspective, we demonstrate the effectiveness of contemporary ML-based AI in augmenting the work of human workers (Raisch and Krakowski 2021). Our AI algorithm is trained solely on past work records without human intervention, a defining feature of contemporary AI. Our findings should help facilitate the wider adoption and usage of this form of AI. In recent years, companies have accumulated large amounts of data through rapid digitization of transactions and can, therefore, potentially realize value through adopting such AI algorithms.

Fourth, this study provides important lessons for successful implementation of AI. A growing number of studies show the potential of AI in complementing human expertise (Fügener et al. 2021, Teodorescu et al. 2021). For instance, van den Broek et al. (2021) depict a hybrid practice that combines AI and human expertise, whereas Sturm et al. (2021) document that AI can collaborate with human learning and reduce a human’s explorative learning for new ideas. A common finding in the literature is that AI is generally less trusted than human experts (Dietvorst et al. 2015); Bogert et al. (2021) find that humans disregard inaccurate advice from AI more strongly than similar advice when it comes from human peers. Most current designs of AI do not incorporate variations among human users (Chakraborti et al. 2017, Carroll et al. 2019, Johnson and Vera 2019, Zhang et al. 2020), which is likely to be a critical issue in real business practice. As Lebovitz (2019) shows, AI could increase ambiguity, so human workers would need to pay an extra cost to collaborate. We show that although senior workers’ experience puts them in a better position to enjoy higher benefits from teaming with AI than their junior colleagues, low trust stymies the realization of significant benefits from the AI.

Finally, our findings indicate certain necessary cautions for managers when implementing AI in work settings. Different worker experience levels should be considered when evaluating job performance in roles that require teaming with AI. New workers with less task experience may be naturally disadvantaged in leveraging AI for high productivity. This leads to the question of how to adjust performance evaluation systems after AI is adopted. Failure to consider the experience difference in evaluating workers’ productivity with AI use may result in unintended discrimination against new workers.

We acknowledge several limitations of this study. The sample is drawn from one company, which may raise concerns about the generalizability of our findings. However, the coding task performed by this company’s employees is typical of knowledge work tasks. In addition, the AI created for this study is representative of AI

at the present developmental stage. Therefore, the conclusions we derive from this study can inform a broader spectrum of contexts in which AI is used. Furthermore, the coding industry is characterized by a high turnover rate. To ensure the continuity of productivity before and after AI adoption, coders in the treatment group were not randomly selected but drawn from the company's more stable workers. We conducted matching and extensive checks (including the falsification test, the placebo test, and verification of a parallel trend) to strengthen the robustness of our findings. One interesting avenue for future research would be to examine the effect of AI in a randomized field experiment. Third, the format of human-AI collaboration can vary significantly across different contexts regarding different tasks and different AI implementations. We examined the format where human knowledge workers are downstream of AI's outputs. We believe this is the dominant form of human-AI collaboration given that AI can quickly perform large volumes of work at a lower cost and the relatively higher acceptance of decisions where human workers are at the final confirmation step.

We note that readers should use caution when generalizing our findings to an AI at a different performance level. With the rapidly evolving technological landscape of machine learning, the theoretical arguments and empirical results must be interpreted in light of the current AI frontier and our setting. This study examines AI as a form of augmented intelligence in assisting human experts to make decisions. Our contribution is applicable to similar implementations of AI for human-AI teaming rather than settings where AI is allowed to operate independently. The defining assumptions of this frontier are that AI is not more intelligent or insightful than human experts for the focal knowledge task. It acts as an aid and is purely functional with no self-awareness, and it does not reshape the role of human experts in the organization. If any of the assumptions are challenged by advances in AI, additional factors would need to be considered as the nature of human-AI teaming is reconstructed.

For future research in AI implementation, our work points to several promising opportunities that warrant further examination. There are unanswered questions within the broad issue that we address in this study: a human worker's ability to work with AI. Contemporary AI achieves its unprecedented performance because of its unique learning ability. However, as its application expands, more research is needed on the mechanisms that can enhance user trust in AI. In this regard, potential solutions include provision of evidence for AI's benevolence and education to senior workers. To enable wide-ranging adoption of AI in knowledge work, organizations will need to proactively devise mechanisms to address the trust deficit among senior workers. Although our study is not directly focused on this challenge, our findings suggest that senior workers need to better understand how AI

performs and be more tolerant of its mistakes. It may be the case that building greater transparency and explainability into the AI can partially address the low trust. Communicating data and sharing data about the performance benefits of working with AI are other ways in which senior workers' concerns about negative impacts on the organization could be assuaged. We also expect that the roles of trust-related factors might change over time as AI keeps improving its performance (Ahn et al. 2021) and becomes more human like (Seymour et al. 2021). More field studies as well as laboratory experiments are required to shed light on this critical challenge.

Furthermore, the dynamic interactions between AI and human workers need further exploration. An interesting question that merits further investigation is related to how AI may transform the meaningfulness of work and whether there is variation in this perception across workers with different levels and types of experience. Does AI enrich knowledge work or diminish its significance because of the different nature of human input required and the extent to which AI may guide the decision process? Qualitative studies may help shed light on this question. Furthermore, the framework in this study also offers valuable opportunities for richer understanding of the mechanisms of human-AI complementarities. What unique advantage does human task experience bring to the human-AI teaming synergy? A deeper investigation on the distinct learning mechanisms of human and AI is required for different stages ranging from training to production use and continuous learning/updating. Finally, our findings also point to the necessity of understanding how to design workflows when human workers and AI are both present. In this study, we let human workers be the "supervisors" of AI. However, more collaboration formats can be designed based on the nature of tasks and on the distribution of human resources. Research is needed to uncover the best strategies for embedding AI into human organizational structures.

Endnotes

¹ Brynjolfsson et al. (2019, p. 24) note that "[m]achine learning represents a fundamental change from the first wave of computerization. Historically, most computer programs were created by meticulously codifying human knowledge, mapping inputs to outputs as prescribed by the programmers. In contrast, machine-learning systems use categories of general algorithms (e.g. neural networks) to figure out relevant mappings on their own, typically by being fed very large sample data sets. By using these machine-learning methods that leverage the growth in total data and data-processing resources, machines have made impressive gains in perception and cognition."

² Imagine that two workers, A and B, joined a company together. A kept working on task 1, whereas B rotated across tasks 1, 2, and 3. In this case, A and B have the same tenure experience, but A has more task experience with task 1.

³ Although the presence of bias in AI is not the focus of this paper, we acknowledge that, to the extent that an AI may be trained on data from humans, it may encapsulate biases.

⁴ Deep learning models using Graphics Processing Unit are far more computationally expensive than off-the-shelf SVM models. They also require extensive configuration of the operational processes, which creates additional costs. Because the deep learning models are not perfect either, the SVM model was selected by managers after consideration of the trade-off (i.e., computational efficiency and model performance).

⁵ Although one might prefer using a random sample of coders as the treatment group, some practical concerns preclude us from doing so. There is a relatively high turnover in coding jobs, and there is a concern that the randomization process would enlist coders who might not have sufficient history for the pretrend data or who might drop out soon, which would introduce bias in the analysis of long-term productivity. We, therefore, opt for a more stable treatment group and perform extensive validation tests that are described later.

⁶ Our sample includes a longer pretreatment period to better examine the trends of the treatment and control groups before the AI implementation. In our robustness test, we also use a shorter pretreatment period (the same length as the posttreatment period) and confirm that all the results remain the same (see Table A3 in the online appendix).

⁷ The number of charts per coder in the two study groups is different, which is largely because of the high turnover rate of this job. As previously mentioned, we selected stable coders for the treatment group, resulting in a lower turnover rate and thus, a higher number of charts per coder.

⁸ The difference is because of the selection of the treatment group as described in footnote 4. In the following analyses, we also use matched samples and a series of robustness checks to address the selection concerns.

⁹ The 10-year cutoff for seniority is based on the recommendation of company management. We repeat our analyses using multiple thresholds for robustness.

¹⁰ Job titles that are classified as managers include “lead coding data review consultant,” “clinical communication center nurse,” “clinical abstractor traveler,” “lead remote coding review consultant,” “manager, clinical quality,” “manager, quality assurance,” and “senior business analyst.” Nonmanager job titles are “review consultant,” “contractor,” “quality assurance team member,” “remote review consultant,” and “site review consultant.”

References

- Agrawal AK, Gans JS, Goldfarb A (2021) AI adoption and system-wide change. NBER Working Paper No. 28811, National Bureau of Economic Research, Cambridge, MA.
- Ahn D, Almaatouq A, Gulabani M, Hosanagar K (2021) Will we trust what we don't understand? Impact of model interpretability and outcome feedback on trust in AI. Preprint, submitted November 16, <https://arxiv.org/abs/2111.08222>.
- AI HLEG (2019) Ethics guidelines for trustworthy AI. Report No. B-1049, Brussels.
- Allen R, Choudhury P (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organ. Sci.* 33(1):149–169.
- Appari A, Johnson ME, Anthony DL (2018) Health IT and inappropriate utilization of outpatient imaging: A cross-sectional study of US hospitals. *Internat. J. Medical Informatics* 109:87–95.
- Atasoy H, Chen PY, Ganju K (2017) The spillover effects of health IT investments on regional healthcare costs. *Management Sci.* 64(6):2515–2534.
- Bai B, Dai H, Zhang D, Zhang F, Hu H (2022) The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments. *Manufacturing Service Oper. Management* 24(6):3060–3078.
- Baniecki H, Kretowicz W, Piatyszek P, Wisniewski J, Biecek P (2021) Dalex: Responsible machine learning with interactive explainability and fairness in Python. *J. Machine Learn. Res.* 22(214):1–7.
- Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E (2019) Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. *Proc. Conf. AAAI Artificial Intelligence* 33(01):2429–2437.
- Bhatt U, Andrus M, Weller A, Xiang A (2020) Machine learning explainability for external stakeholders. Preprint, submitted July 10, <https://arxiv.org/abs/2007.05408>.
- Bogert E, Schechter A, Watson RT (2021) Humans rely more on algorithms than social influence as a task becomes more difficult. *Sci. Rep.* 11(1):1–9.
- Borman WC, Hanson MA, Oppler SH, Pulakos ED, White LA (1993) Role of early supervisory experience in supervisor performance. *J. Appl. Psych.* 78(3):443–449.
- Bradley JV (1981) Overconfidence in ignorant experts. *Bull. Psychonomic Soc.* 17(2):82–84.
- Bresnahan T (2021) Artificial intelligence technologies and aggregate growth prospects. Diamond J, Zodrow G, eds. *Prospects for Economic Growth in the United States* (Cambridge University Press, Cambridge, UK), 132–152.
- Bresnahan TF, Brynjolfsson E, Hitt LM (2002) Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quart. J. Econom.* 117(1):339–376.
- Brynjolfsson E (2022) The Turing trap: The promise & peril of human-like artificial intelligence. *Daedalus* 151(2):272–287.
- Brynjolfsson E, Hui X, Liu M (2018a) Does machine translation affect international trade? Evidence from a large digital platform. NBER Working Paper No. 24917, National Bureau of Economic Research, Cambridge, MA.
- Brynjolfsson E, Rock D, Syverson C (2018b) Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. Agrawa A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 23–57.
- Brynjolfsson E, Rock D, Syverson C (2019) Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. Agrawa A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence* (University of Chicago Press, Chicago), 23–60.
- Bundorf MK, Polyakova M, Tai-Seale M (2022) How do humans interact with algorithms? Experimental evidence from health insurance. NBER Working Paper No. 25976, National Bureau of Economic Research, Cambridge, MA.
- Burt A (2019) The AI transparency paradox. *Harvard Bus. Rev.* (December 13), <https://hbr.org/2019/12/the-ai-transparency-paradox>.
- Card D, Krueger AB (1993) Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. NBER Working Paper No. 4509, National Bureau of Economic Research, Cambridge, MA.
- Carroll M, Shah R, Ho MK, Griffiths T, Seshia S, Abbeel P, Dragan A (2019) On the utility of learning about humans for human-AI coordination. *Adv. Neural Inform. Processing Systems*, vol. 32, 5174–5185.
- Case A, Deaton A (2017) Mortality and morbidity in the 21st century. *Brookings Papers Econom. Activity* 2017:397–476.
- Cellan-Jones R (2014) Stephen Hawking warns artificial intelligence could end mankind. BBC News (December 2), <https://www.bbc.com/news/technology-30290540>.
- Chakraborti T, Kambhampati S, Scheutz M, Zhang Y (2017) AI challenges in human-robot cognitive teaming. Preprint, submitted July 15, <https://arxiv.org/abs/1707.04775>.
- Clement J, Ren Y, Curley S (2020) Disregarding, modifying or adopting: How medical experts incorporate AI recommendations into patient care decisions. Agarwal R, Gao GG, Crowley K, McCullough J, eds. *Conf. Health IT Analytics (CHITA)*, 32.
- Crawford M (2013) The truth about computer-assisted coding. *J. AHIMA* 84(7):24–27.

- Dai T, Singh S (2020) Conspicuous by its absence: Diagnostic expert testing under uncertainty. *Marketing Sci.* 39(3):540–563.
- Dai T, Singh S (2023) Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. Preprint, submitted January 13, <http://dx.doi.org/10.2139/ssrn.3987454>.
- Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. *Proc. Natl. Acad. Sci. USA* 108(17):6889–6892.
- Davenport TH, Ronanki R (2018) Artificial intelligence for the real world. *Harvard Bus. Rev.* 96(1):108–116.
- David H (2015) Why are there still so many jobs? The history and future of workplace automation. *J. Econom. Perspect.* 29(3):3–30.
- Davis E, Marcus G (2015) Commonsense reasoning and commonsense knowledge in artificial intelligence. *Comm. ACM* 58(9):92–103.
- De Martino B, Kumaran D, Seymour B, Dolan RJ (2006) Frames, biases, and rational decision-making in the human brain. *Science* 313(5787):684–687.
- Decker M, Fischer M, Ott I (2017) Service robotics and human labor: A first technology assessment of substitution and cooperation. *Robotics Autonomous Systems* 87:348–354.
- Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. General* 144(1):114–126.
- Dimick C (2010) Achieving coding consistency. *J. AHIMA* 81(7):24–28.
- Dinan E, Logacheva V, Malykh V, Miller A, Shuster K, Urbanek J, Kiela D, et al. (2020) The second conversational intelligence challenge (convai2). *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations* (Springer International Publishing, Berlin), 187–208.
- Dong W, Li J, Yao R, Li C, Yuan T, Wang L (2016) Characterizing driving styles with deep learning. Preprint, submitted October 8, <https://arxiv.org/abs/1607.03611>.
- Dunlosky J, Hertzog C (2000) Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psych. Aging* 15(3):462–474.
- Ekins S (2016) The next era: Deep learning in pharmaceutical research. *Pharmaceutical Res.* 33(11):2594–2603.
- Fjelland R (2020) Why general artificial intelligence will not be realized. *Humanities Soc. Sci. Comm.* 7(1):1–9.
- Ford JK, Sego D, Quinones M, Speer J (1991) The construct of experience: A review of the literature and needed research directions. *Sixth Annual Conf. Soc. Indust. Organ. Psych.* (Society for Industrial and Organizational Psychology, Columbus, OH).
- Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Quart.* 45(3):1527–1556.
- Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human-artificial intelligence collaboration: Investigating the path toward productive delegation. *Inform. Systems Res.* 33(2):678–696.
- Ganju KK, Pavlou PA, Banker RD (2016) Does information and communication technology lead to the well-being of nations? A country-level empirical investigation. *MIS Quart.* 40(2):417–430.
- Ge R, Zheng Z, Tian X, Liao L (2021) Human-robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Inform. Systems Res.* 32(3):774–785.
- Goldin C, Katz LF (1998) The origins of technology-skill complementarity. *Quart. J. Econom.* 113(3):693–732.
- Goodman PS, Leyden DP (1991) Familiarity and group productivity. *J. Appl. Psych.* 76(4):578–586.
- Grand View Research (2018) Medical coding market size, share & trends analysis report by classification system (International Classification of Diseases, Healthcare Common Procedure Code System), by component, and segment forecasts, 2018–2025. Report, Grand View Research, San Francisco.
- Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316(22):2402–2410.
- Guo Y, Yin X, Liu D, Xu SX (2020) “She is not just a computer”: Gender role of AI chatbots in debt collection. *Internat. Conf. Inform. Systems 2020 (ICIS)* (IEEE, Piscataway, NJ).
- Hainc N, Federau C, Stieltjes B, Biatow M, Bink A, Stippich C (2017) The bright, artificial intelligence-augmented future of neuroimaging reading. *Frontiers Neurology* 8:489.
- Hayes-Roth F, Waterman DA, Lenat DB (1983) *Building Expert Systems* (Addison-Wesley Longman Publishing Co., Inc., Boston).
- He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on I mageNet classification. *Proc. IEEE Internat. Conf. Comput. Vision* (IEEE, Piscataway, NJ), 1026–1034.
- Henry-Nickie M (2017) AI should worry skilled knowledge workers too. *Brookings.edu* (November 8), <https://www.brookings.edu/blog/techtank/2017/11/08/ai-should-worry-skilled-knowledge-workers-too/>.
- Hosnagar K (2019) *A Human's Guide to Machine Intelligence: How Algorithms Are Shaping Our Lives and How We Can Stay in Control* (Viking, New York).
- Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ (2018) Artificial intelligence in radiology. *Nature Rev. Cancer* 18(8):500–510.
- Huang L, Bras RL, Bhagavatula C, Choi Y (2019) Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. *Proc. 2019 Conf. Empirical Methods Natural Language Processing and the 9th Internat. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics), 2391–2401.
- Jabbour S, Fouhey D, Kazerooni E, Sjoding MW, Wiens J (2020) Deep learning applied to chest X-rays: Exploiting and preventing shortcuts. Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, Wiens J, eds. *Machine Learn. Healthcare Conf.* (PMLR, New York), 750–782.
- Jajal TD (2018) Distinguishing between narrow AI, general AI and super AI. *Medium* (May 21), <https://medium.com/mapping-out-2050/distinguishing-between-narrow-ai-general-ai-and-super-ai-a4bc44172e22>.
- Jha S, Topol EJ (2016) Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA* 316(22):2353–2354.
- Johnson M, Vera A (2019) No AI is an island: The case for teaming intelligence. *AI Magazine* 40(1):16–28.
- Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245):255–260.
- Judge TA (1994) Person-organization fit and the theory of work adjustment: Implications for satisfaction, tenure, and career success. *J. Vocational Behav.* 44(1):32–54.
- Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inform. Systems Res.* 32(3):713–735.
- Kahneman D (2011) *Thinking, Fast and Slow* (Macmillan, New York).
- Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.
- Karahanna E, Chen A, Liu QB, Serrano C (2019) Capitalizing on health information technology to enable advantage in us hospitals. *MIS Quart.* 43(1):113–140.
- Klein KJ, Dansereau F, Hall RJ (1994) Levels issues in theory development, data collection, and analysis. *Acad. Management Rev.* 19(2):195–229.
- Knight W (2017) The dark secret at the heart of AI. *MIT Tech. Rev.* (April 11), <https://www.technologyreview.com/2017/04/11/51113/the-dark-secret-at-the-heart-of-ai/>.
- Kohn ML, Schooler C (1982) Job conditions and personality: A longitudinal assessment of their reciprocal effects. *Amer. J. Sociol.* 87(6):1257–1286.

- Korinek A, Stiglitz JE (2017) Artificial intelligence and its implications for income distribution and unemployment. NBER Working Paper No. 24174, National Bureau of Economic Research, Cambridge, MA.
- Kose I, Gokturk M, Kilic K (2015) An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Appl. Soft Comput.* 36(C):283–299.
- Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T (2017) Artificial intelligence in precision cardiovascular medicine. *J. Amer. College Cardiology* 69(21):2657–2664.
- Krusell P, Ohanian LE, Ríos-Rull JV, Violante GL (2000) Capital-skill complementarity and inequality: A macroeconomic analysis. *Econometrica* 68(5):1029–1053.
- Lance CE, Hedge JW, Alley WE (1989) Joint relationships of task proficiency with aptitude, experience, and task difficulty: A cross-level, interactional study. *Human Performance* 2(4):249–272.
- Lebovitz S (2019) Diagnostic doubt and artificial intelligence: An inductive field study of radiology work. *ICIS 2019 Proc.* (IEEE, Piscataway, NJ).
- Lebovitz S, Levina L, Lifshitz-Assa H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quart.* 45(3):1501–1526.
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.
- Li P, Kim MM, Doshi JA (2010) Comparison of the performance of the CMS Hierarchical Condition Category (CMS-HCC) risk adjuster with the Charlson and Elixhauser comorbidity measures in predicting mortality. *BMC Health Services Res.* 10(1):245.
- Li X, Hess TJ, Valacich JS (2008) Why do we trust new technology? A study of initial trust formation with organizational information systems. *J. Strategic Inform. Systems* 17(1):39–71.
- Liu X, Stoutenborough J, Vedlitz A (2017) Bureaucratic expertise, overconfidence, and policy choice. *Governance (Oxford)* 30(4):705–725.
- Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *J. Consumer Res.* 46(4):629–650.
- Lou B, Wu L (2021) AI on drugs: Can artificial intelligence accelerate drug development? Evidence from a large-scale examination of bio-pharma firms. *MIS Quart.* 45(3):1451–1482.
- Lum K (2017) Limitations of mitigating judicial bias with machine learning. *Nature Human Behav.* 1(7):0141.
- Luo X, Qin MS, Fang Z, Qu Z (2021) Artificial intelligence coaches for sales agents: Caveats and solutions. *J. Marketing* 85(2):14–32.
- Manyika J, Lund S, Chui M, Bughin J, Woetzel J, Batra P, Ko R, Sanghvi S (2017) Jobs lost, jobs gained: Workforce transitions in a time of automation. *McKinsey Global Inst.* (November 28), <https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>.
- Maruthappu M, Gilbert BJ, El-Harasis MA, Nagendran M, McCulloch P, Duclos A, Carty MJ (2015) The influence of volume and experience on individual surgical performance: A systematic review. *Ann. Surgery* 261(4):642–647.
- Mayer RC, Davis JH, Schoorman FD (1995) An integrative model of organizational trust. *Acad. Management Rev.* 20(3):709–734.
- McCarthy J (2007) From here to human-level AI. *Artificial Intelligence* 171(18):1174–1182.
- McCauley CD, Ruderman MN, Ohlott PJ, Morrow JE (1994) Assessing the developmental components of managerial jobs. *J. Appl. Psych.* 79(4):544–560.
- McCradden MD, Joshi S, Mazwi M, Anderson JA (2020) Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digital Health* 2(5):e221–e223.
- McDaniel MA, Schmidt FL, Hunter JE (1988) Job experience correlates of job performance. *J. Appl. Psych.* 73(2):327.
- McEnroe MP (1988) Length of experience and the performance of managers in the establishment phase of their careers. *Acad. Management J.* 31(1):175–185.
- McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788):89–94.
- McKnight DH, Choudhury V, Kacmar C (2002) Developing and validating trust measures for e-commerce: An integrative typology. *Inform. Systems Res.* 13(3):334–359.
- McKnight DH, Carter M, Thatcher JB, Clay PF (2011) Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Management Inform. Systems* 2(2):1–25.
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.
- Mortimer JT, Lorence J (1979) Work experience and occupational value socialization: A longitudinal study. *Amer. J. Sociol.* 84(6):1361–1385.
- Ng TW, Feldman DC (2010) Organizational tenure and job performance. *J. Management* 36(5):1220–1250.
- Nikolaidis S, Shah J (2013) Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. 2013 8th ACM/IEEE Internat. Conf. Human-Robot Interaction (HRI) (IEEE, Piscataway, NJ), 33–40.
- Ostroff C, Ford JK (1989) Assessing training needs: Critical levels of analysis. Goldstein IL, ed. *Training and Development in Organizations* (Jossey-Bass, San Francisco), 25–62.
- Pakdemirli E (2019) Artificial intelligence in radiology: Friend or foe? Where are we now and where are we heading? *Acta Radiologica Open* 8(2):2058460119830222.
- Pasquale F (2015) *The Black Box Society* (Harvard University Press, Cambridge, MA).
- Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Iezzoni LI, Ingber MJ, Levy JM, Robst J (2004) Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financial Rev.* 25(4):119–141.
- Puig X, Shu T, Li S, Wang Z, Liao YH, Tenenbaum JB, Fidler S, Torralba A (2020) Watch-and-help: A challenge for social perception and human-AI collaboration. Preprint, submitted May 3, <https://arxiv.org/abs/2010.09890>.
- PwC (2017) Sizing the prize: What's the real value of AI for your business and how can you capitalise? Report. <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>.
- Quinones MA, Ford JK, Teachout MS (1995) The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personality Psych.* 48(4):887–910.
- Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation-augmentation paradox. *Acad. Management Rev.* 46(1):192–210.
- Ribeiro MT, Singh S, Guestrin C (2016) “Why should I trust you?” Explaining the predictions of any classifier. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 1135–1144.
- Ricard S (2020) The year of the knowledge worker. *Forbes* (December 10), <https://www.forbes.com/sites/forbestechcouncil/2020/12/10/the-year-of-the-knowledge-worker/?sh=1fdb242d7fbb>.
- Rossi F (2018) Building trust in artificial intelligence. *J. Internat. Affairs* 72(1):127–134.
- Rudin P (2017) Thoughts on human learning vs. machine learning. *Singularity2030.ch* (January 13), <https://singularity2030.ch/thoughts-on-human-learning-vs-machine-learning/>.
- Sadigh D, Landolfi N, Sastry SS, Seshia SA, Dragan AD (2018) Planning for cars that coordinate with people: Leveraging effects on

- human actions for planning and active information gathering over human internal state. *Autonomous Robots* 42(7):1405–1426.
- Sarkar D (2018) The importance of human interpretable machine learning. *Medium* (May 24), <https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476>.
- Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *J. Amer. Medical Inform. Assoc.* 17(5): 507–513.
- Schmidt FL, Hunter JE, Outerbridge AN (1986) Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *J. Appl. Psych.* 71(3): 432–439.
- Seymour M, Kriemer K, Yuan L, Dennis A (2021) Beyond deep fakes: Conceptual framework and research agenda for neural rendering of realistic digital faces. *Proc. 54th Hawaii Internat. Conf. System Sci. HICSS* (IEEE, Piscataway, NJ), 4859–4868.
- Siu HC, Pena JD, Chen E, Zhou Y, Lopez VJ, Palko K, Chang KC, Allen RE (2021) Evaluation of human-AI teams for learned and rule-based agents in Hanabi. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Conf. Neural Inform. Processing Systems (NeurIPS)*, vol. 34 (NeurIPS, San Diego), 16183–16195.
- Slocum JW Jr, Cron WL, Hansen RW, Rawlings S (1985) Business strategy and the management of plateaued employees. *Acad. Management J.* 28(1):133–154.
- Sturm T, Gerlach JP, Pumplun L, Mesbah N, Peters F, Tauchert C, Nan N, Buxmann P (2021) Coordinating human and machine learning for effective organizational learning. *MIS Quart.* 45(3): 1581–1602.
- Sun C, Shi ZJ, Liu X, Ghose A, Li X, Xiong F (2021) The effect of voice AI on consumer purchase and search behavior. Working paper, NYU Stern School of Business, New York.
- Swap W, Leonard D (2014) Artificial intelligence can't replace hard-earned knowledge—Yet. *Harvard Bus. Rev.* (November 17), <https://hbr.org/2014/11/artificial-intelligence-cant-replace-hard-earned-knowledge-yet>.
- Syverson C (2017) Challenges to mismeasurement explanations for the US productivity slowdown. *J. Econom. Perspect.* 31(2):165–186.
- Taddy M (2018) The technological elements of artificial intelligence. Agrawal A, Gans J, Goldfarb A, ed. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 61–88.
- Talley EM, Newman D, Mimno D, Herr BW II, Wallach HM, Burns GA, Leenders AM, McCallum A (2011) Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods* 8(6):443–444.
- Tan TF, Netessine S (2020) At your service on the table: Impact of tabletop technology on restaurant performance. *Management Sci.* 66(10):4496–4515.
- Tandon N, Varde AS, de Melo G (2018) Commonsense knowledge in machine intelligence. *SIGMOD Record* 46(4):49–52.
- Teodorescu MH, Morse L, Awwad Y, Kane GC (2021) Failures of fairness in automation require a deeper understanding of human-ML augmentation. *MIS Quart.* 45(3):1483–1500.
- Tesluk PE, Jacobs RR (1998) Toward an integrated model of work experience. *Personality Psych.* 51(2):321–355.
- Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.
- Tzeng OJ (1973) Positive recency effect in a delayed free recall. *J. Verbal Learn. Verbal Behav.* 12(4):436–439.
- Vance RJ, Coover MD, MacCallum RC, Hedge JW (1989) Construct models of task performance. *J. Appl. Psych.* 74(3):447–455.
- van den Broek E, Sergeeva A, Huysman M (2021) WHEN the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quart.* 45(3):1557–1580.
- Veiga JF (1981) Plateaued versus nonplateaued managers: Career patterns, attitudes, and path potential. *Acad. Management J.* 24(3): 566–578.
- Wagner JA III, Ferris GR, Fandt PM, Wayne SJ (1987) The organizational tenure—Job involvement relationship: A job-career experience explanation. *J. Organ. Behav.* 8(1):63–70.
- Welleck S, Weston J, Szlam A, Cho K (2018) Dialogue natural language inference. Preprint, submitted January 18, <https://arxiv.org/abs/1811.00671>.
- Williams C (2015) AI guru Ng: Fearing a rise of killer robots is like worrying about overpopulation on Mars. *The Register*, https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai/.
- Wu L, Kane GC (2021) Network-biased technical change: How modern digital collaboration tools overcome some biases but exacerbate others. *Organ. Sci.* 32(2):273–292.
- Zhang R, McNeese NJ, Freeman G, Musick G (2020) 'An ideal human' expectations of AI teammates in human-AI teaming. *Proc. ACM Human-Comput. Interaction* 4(CSCW3) (ACM, New York), 1–25.
- Zysman J, Nitzberg M (2020) Governing AI: Understanding the limits, possibility, and risks of AI in an era of intelligent tools and systems. Preprint, submitted October 14, <http://dx.doi.org/10.2139/ssrn.3681088>.