Decoupled learning among medical trainees and artificial intelligence

Bryan Spencer
Frankfurt School of Finance & Management
Adickesallee 32-34
Frankfurt am Main, D-60322
Germany
Email: b.spencer@fs.de

Abstract

One of the great paradoxes of AI development is that even when models perform well in testing, their real world performance often falls short. Existing research elides the complexities of learning processes in the "human-algorithm" interaction during the development of artificial intelligence, despite its importance in shaping outcomes. In order to study these learning processes, I conducted a 13-month ethnographic study of an elite teaching hospital, "Southern Eye Hospital," one of the first hospitals in China to establish an AI Department, where medical trainees learned alongside algorithms through repetitive enactments of "labeling routines." I illustrate how the development of artificial intelligence has implications for learning with "decoupled learning," which emphasizes how certain forms of expertise achieved by trainees and artificial intelligence models may be too abstract to transfer into real world application.

> *"I think if you work as a radiologist, you're like the coyote that's already over the edge of the cliff but hasn't yet looked down … People should stop training radiologists now, it's just completely obvious that within five years, deep learning is going to do better than radiologists"* (Hinton, 2016). Geoff Hinton, Turing Award recipient for contributions to deep learning.

One of the major paradoxes of artificial intelligence (AI) development in medicine is that even when models developed using machine learning algorithms perform well on paper, real-world performance often fails to live up to expectations (Topol, 2020). Social scientists have asserted that a variety of factors may explain this discrepancy, such as failure to incorporate social and tacit knowledge into AI tools (Lebovitz, Levina, & Lifshitz-Assaf, 2021) or unanticipated environmental factors that manifest during AI tool use (Beede et al., 2020). Yet these explanations are derived from post-hoc evaluations of the AI tools during use. They fail to capture the learning processes of "human-algorithm" interaction during AI development, where the divergence between "expected" and "real" performance comes apart. As questions about the future of professional expertise in the age of AI remain unanswered (Faraj, Pachidi, & Sayegh, 2018; Kellogg, Valentine, & Christin, 2020), turning attention to the development of AI offers a path to push scholarship and society forward (Bailey & Barley, 2020).

The development of AI tools using retrospective data (i.e., machine learning algorithms trained with pre-existing "big data" such as medical records, images, etc. to create an AI model) can be understood through a situated learning perspective, which shows that learning occurs when more experienced members pass down knowledge to newcomers (Lave & Wenger, 1991), who learn in the process of "doing" (Brown & Duguid, 1991; Eisenhardt & Tabrizi, 1995). AI models are trained using large data sets that are collected, cleaned, standardized, and labeled by experts. Labeling, also referred to as annotation, is a computer-mediated routine enactment, or "repetitive, recognizable pattern of interdependent actions, involving multiple actors" (Feldman & Pentland, 2003: 95), in which actors categorize and codify data that can be used to train machine learning algorithms (Brynjolfsson & Mitchell, 2017). For example, in medicine, labeling routines may entail classifying and diagnosing diseases using clinical data.

Recent studies have looked at how experienced professionals are tasked with taking their skills and knowledge and applying it to improving data sets (Sachs, 2020). For example, Pine and Bossen's (2020) study of "Clinical Documentation Integrity Specialists" examined how Registered Nurses, all with several years of experience, worked to create structured language for

administrative documentation datasets in hospitals (see also, Pine et al., 2021). Lebovtiz, Levina, and Lifshitz-Assaf's (2021) studied five different AI tools found that "ground truth" labels were developed by experienced physicians. These studies approach labeling routines with the assumption that the professionals involved already have the preexisting expertise needed for the routine performances, despite evidence to the contrary (e.g., Gray & Suri, 2019; Metz, 2019). In scenarios where preexisting expertise does not exist, new entrants to a profession learn alongside machine learning algorithms during the labeling routine enactment. No studies, to my knowledge, have examined what happens when new entrants to a profession are tasked with labeling routines (i.e., labeling data for AI). In part, the paucity of studies in this space may be explained by the difficulty of getting access field sites where AI is developed (Christin, 2020). To achieve expert levels, new entrants need to perform the labeling routine thousands of times. Labeling routines are an integral part of the AI development process (Deng et al., 2009), and bring to the fore the importance of understanding the relationship between cognition and routine enactments (Lazaric, 2008; 2021), which has largely been limited to experimental and simulation work in the past (e.g., Cohen & Bacdayan, 1994; Miller, Pentland, & Choi, 2012). This leads to a question of how repetitive enactments of "labeling routines" influence the way new entrants into a profession learn to become experts, and what are the consequences of this learning process?

I study this question in the context of labeling routines that took place at "Southern Eye Hospital" (pseudonym), an elite teaching hospital in China specializing in ophthalmology. Southern was one of the first hospital's in China to establish an AI Department, where a team of doctors and trainees (postgraduate medical students and residents) worked with a technical team to develop AI tools. While subject matter experts such as senior doctors seem to be the logical choice to annotate data for AI models, the scope of projects, where hundreds of thousands of images, medical records, etc. must be sorted and labeled, meant that senior doctors could not commit their time to this work, and instead assigned the task to trainees. To enact labeling routines, trainees had to learn diagnostic skills that would allow them to achieve "expert levels" in their labeling accuracy. I found that trainees' accuracy labeling images increased throughout routine enactments, reaching levels equivalent to specialists. Further, the AI models they created performed at high levels, with accuracy comparable to that of specialists. Yet, the development of the models also led to an apparent paradox. While senior doctors were able to "translate" their diagnostic skills into teaching trainees how to label images (a task many senior doctors have

never done before), the reverse was less straightforward. Despite the high levels of accuracy indicated by the trainees and algorithms alike in the model performance, both struggled to translate that into clinical practice. Trainees, despite being experts at labeling images of diseases, were unable to diagnose the same diseases when facing 'real' people in the clinic. Further, most doctors felt that the AI, while accurate in testing, was not useful in clinical work. The development of the trainees and the AI model seemed to parallel each other, both in terms of diagnostic accuracy, but also limited clinical ability.

In this study, I develop a theoretical model that demonstrates how repetitive, computer-mediated learning can disrupt the mutual constitution of performing and patterning in routines. This suggests that much like an AI model can be trained to an excessive degree and lead to "overfitting" (in which an AI model struggles to generalize to new data), trainees that engage in extensive repetitions of labeling routines struggle to generalize the skills acquired from labeling routines when enacting diagnosis routines. Theoretically, the routine dynamics lens helps unpack the cognitive, physical, and emotional aspects of situated learning in the labeling and diagnosis routine enactments, while also resolving the "expertise paradox" I observed in the field.

SITUATED LEARNING IN COMMUNITIES OF PRACTICE AND ROUTINE DYNAMICS

Since seminal work by Lave and Wenger in 1991, scholars of learning in communities of practice have primarily focused on situated learning as a "socialization" process (Wenger, 1999), which focuses on how newcomers learn through participation in communities. For example, Cook and Brown (1999) chronicled how apprentice flutemakers learned to craft world-class flutes through iterative interactions with master flutemakers. Master flutemakers' feedback centered around the "feel" of flutes, emphasizing tacit knowledge that could only be gained through the process of repeatedly crafting flutes. Cook and Brown's example of apprentice flutemakers "learning by doing" represents a core concept of legitimate peripheral participation, a process in which trainees participate in the practice of an expert to a limited degree in order to learn "knowledgeable skills" within a community of practice (Lave & Wenger, 1991: 29).

Scholars of learning in communities of practice have studied the use of "abstract rehearsals" (Beane, 2019) in training professionals. Abstract rehearsals are simulations that represent a "real world" work. In medicine, there are broad applications of abstract rehearsals as a means through which to train new members of the profession (Bailey, Leonardi, & Barley,

2012), such as surgeons training on simulators (Beane, 2019; Johnson, 2007) or "re-assimilating" knowledge of anatomical images to that of the patient's body (Hirschauer, 1991; Koschmann et al., 2011). These representations are central to the concept of legitimate peripheral participation.

Recent work examining simulations has rekindled the notion of linking the cognitive and social processes of learning. For example, Beane's (2019) study demonstrated how surgical trainees gained robotic surgery skills through training on software simulators that were attached to surgical consoles. Through this "shadow learning," trainees developed the skills needed when attending physicians offered the opportunities to take over in the operating room. Bridging the cognitive and situated aspects of learning seems of particular importance when considering the increasing use of technological simulations in training new members of professions (Bailey, Leonardi, & Barley, 2012). Consider, for example, that simulations can be either physical, virtual, or both (Prentice, 2005). Simulations, particularly virtual ones, can increase cognitive representations of learning while downplaying the situated, performative aspects of the "body" (Bailey, Leonardi, & Barley, 2012).

When new technologies are introduced into a workplace, learning is often facilitated through routines (Edmondson, Bohmer, & Pisano, 2001; Kellogg et al. 2021). Like studies of situated learning, which view learning as "local and composed of artifacts, people, and tasks" (Bailey and Barley, 2011: 264), routine dynamics examines situated, local practices, tracing how actions, actors, and artifacts involved in routine enactments evolve over time (Feldman et al., 2021). Routine dynamics scholarship traces its history back to the Carnegie School (Cyert & March, 1963; March & Simon, 1958; Simon, 1947; Nelson & Winter, 1982) where routines were seen as largely cognitive patterns that influenced organizational behavior, manifesting through standard operating procedures. In this literature, routines were a source of "cognitive efficiency," or reducing complexity of tasks by "routinizing" them (Salvato & Rerup, 2011).

Feldman and Pentland (2003: 95) argued that organizational routines, in addition to having a cognitive aspect, "consist of the resulting performances and the understandings of these performances." Over the past two decades, the cognitive aspects of routines have been backgrounded (Lazaric, 2021; Levinthal & Rerup, 2006; Rerup & Spencer, 2021) as routine dynamics scholars shifted towards practice theory and embraced an ontological approach that saw routine enactments as consisting of both 'patterning' and 'performing' elements (Feldman, 2016). Patterning is the abstract model of understanding and enacting a routine. Performing is the

"specific actions by specific people at specific times and places that bring the routine to life" (Feldman & Pentland, 2003: 94). This action-centric view suggests that patterns of routines are produced as routines are performed (Pentland & Goh, 2019). In other words, patterning and performing (of routines) are mutually constituted (Feldman, 2016: 38).

Much like how scholars of learning in communities of practice pushed away from cognitive patterns towards studying the "doing" or social aspects of learning in recent years, scholarship on routines over the last two decades has emphasized the "doing" or performative aspects of routine enactments. As a result, little is known about how "cognitive constructs influence performing and patterning of routines" (Rerup & Spencer, 2021: 454). This is problematic because computer-mediated routine enactments may make the physical, performative aspects of routine performances difficult to detect or non-existent (Barley & Kunda, 2001). In virtual simulators, for example, learning (and routine enactments) may largely be cognitive: learning in these instances would manifest as a process of patterning (Rerup & Spencer, 2021). For labeling routines, the cognitive processes that constitute learning may be become decoupled or segmented from the physical enactment of performing routines. For routine dynamics scholars, the implications of this "virtual shift" of labeling routines would challenge the notion that learning is based on the outcome of routine performances involving multiple actors (Feldman & Pentland, 2003; Miller, Pentland, & Choi, 2012). Bailey, Leonardi, and Barley (2012: 1489) note one way in which problems with learning can manifest in virtual forms:

> "When workers substitute digital simulations for objects or people, they no longer simply operate with, on, or even through representations. They begin to *operate within* them. Operating within a representation means that the worker's connection to the referent is suspended."

When learners are immersed in a purely virtual environment, they risk getting lost within them. Trainees in Beane's (2019: 105) study were able to avoid this as simulators required a performative aspect in which "trainees had to master a new bodily grammar for their professional work." Machine learning algorithms, on the other hand, require thousands of repetitions of computer-mediated labeling routine enactments, far from the diagnosis routine performances in clinics that the patterning of labeling routines were initially based on. As a result, the repetitive,

computer-mediated forms of learning may drive a wedge between the mutual constitution of performing and patterning routines, the implications of which are unknown.

METHODS

Research Context: AI Development at Southern Eye Hospital

Southern was one of the first hospitals in China to establish an AI Department, where doctors and medical trainees work alongside a technical team of developers and engineers to develop, test, and deploy medical AI tools. As mentioned previously, the development of AI tools entails collecting, cleaning, standardizing, and labeling data. AI tools could either be developed "prospectively" or "retrospectively." Prospective AI studies are first designed, then patients are recruited and data is collected. Retrospective AI studies, on the other hand, utilize pre-existing "big data," such as medical records or images. Southern treated more than one million patients per year, meaning they had large amounts of pre-existing data from which to develop retrospective AI projects. In this paper, I focus on the development of nine retrospective AI projects at Southern: Amaranth, Ecru, Cordovan, Glaucous, Icterine, Lanzones, Malachite, Olivine, Sienna (pseudonyms). These projects all utilized pre-existing data from Southern (sometimes supplemented with data from partner hospitals). The goal of each of these projects was to develop an AI tool that could detect one or more eye-related diseases.

Trainees described data prep (cleaning and standardizing retrospective data) and labeling as a "dirty job," akin to scut work (Huising, 2015). To successfully develop AI tools, trainees needed to diagnose and label data at an equivalent level to specialists, otherwise the AI models would not perform well. Labeled data is used to train machine learning algorithms in an iterative process. Later, validation and test data are used to finetune the model and validate its performance (Beam and Kohane, 2016). The result is a machine learning model (or "AI tool") which is able to recognize objects or patterns based on the training data (Faraj, Pachidi, and Sayegh, 2018).

Data Collection

I conducted a 13-month ethnographic study of Southern between September 2019 and September 2020[1]. I negotiated full access to the hospital and conducted fieldwork five to six days a week. In total, I spent 1,043 hours in the field conducting observations. For retrospective

[1] I was out of the field from the end of January until March 1st (five weeks total) due to COVID-19, along with all personnel from the hospital. I participated in daily group meetings online to keep connected to informants.

projects, I sat with trainees as they enacted labeling routines, annotating medical records and images in the AI Department's office (located on an upper floor above the surgical suites and clinical rooms where patients were). In each case, I took notes on actions, dialogues, and interactions as they went through the AI development process. At the beginning of the study, I spent extensive time shadowing and interviewing doctors and trainees in "regular" (non-AI) clinical work to understand what their typical jobs were like.

Additionally, I conducted a total of 70 interviews (ranging from 30 to 125 minutes each) with 41 different members of Southern. Interviews were recorded and transcribed. 16 of the interviews were with doctors and trainees that were not involved with AI. For example, I spoke with trainees that focused on clinical (non-AI) work to develop an understanding of diagnosis routine enactments in the clinics. I used this to develop a comparative understanding of the labeling routine enactments by trainees in the AI Department. In interviews with those in the AI Department, I included questions on 1) their role and tasks during AI projects 2) their prior experience and knowledge of AI and medicine (e.g., knowledge about specific diseases they worked on for AI tools), 3) challenges that the trainees faced during development and why they thought these occurred and 4) how clinical experiences, such as diagnosing patients, compared to labeling routine enactments. In addition to semi-structured interviews, I made use "real-time interviews" (Barley and Kunda, 2001: 85) when trainees were enacting computer-mediated labeling routines. I would ask questions while observing their actions, or ask them to walk me through the process of a labeling routine enactment so that I could understand what was happening on the computer.

To supplement observations and interviews, I collected and analyzed 813 documents that were shared internally on AI projects. For example, documents included PowerPoint slides from sessions instructing trainees how to recognize various eye-related diseases for labeling routines. I also was provided with screenshots of conversations from chat groups related to these projects—where students asked about difficult-to-diagnose images during AI development. Lastly, I captured 2,129 photos and 134 videos of life in the hospital.

Data Analysis

My analysis draws primarily upon data collected around nine retrospective AI projects being developed at Southern. Initially, I constructed timelines of each of the projects. Drawing on interviews, observations, and archival data, I identified phases in the AI development where

labeling was the primary activity. Creating these "temporal brackets" (Langley, 1999) helped me trace the actors, actions and artifacts involved in the labeling routine enactments. Further, this bracketing strategy was using for tracing learning during routine enactments, as well as theorizing on the outcomes that followed.

Routine dynamics was useful as an analytical lens (Feldman et al., 2021; Parmigiani & Howard-Grenville, 2011) to understand the learning that unfolded in the labeling routine performances. A routine dynamics perspective captured the iterative work that goes into labeling data and developing AI: Senior doctors taught trainees how to diagnose and label data. Trainees, in turn, taught undergraduate students. This iterative process repeated for thousands of pieces of data as a pattern of interdependent actions (learning to label data, labeling data, and iterating in the process of training machine learning algorithms) involving multiple actors (senior doctors, trainees, and undergraduate students).

During data collection in the field, I wrote reflective memos based on emerging themes that I was observing. While in many studies of situated learning, "measuring learning outcomes [is] difficult" (Kellogg et al., 2021: 188), in AI development there are clear, measurable outcomes of learning. I traced learning by routine participants across different levels of the professional hierarchy. For example, trainees leading the projects were graduate students that were in the AI Department in lieu of clinical rotations in other "traditional" departments. To lead AI projects, they had to learn from sources other than the clinic. Undergraduate students hired to annotate were at earlier stages of their education, and relied on trainees to learn how to enact labeling routines.

My analytic strategy was inductive and based on grounded theory (Glaser & Strauss, 1967). Before coding my data, I read my field notes and interview transcripts multiple times. My initial coding focused on actions and interactions (Strauss, 1978) of routine participants. To this end, I used "process coding," assigning labels that captured the "action" in the data (Saldaña, 2014). I engaged in constant comparisons of my data (Strauss & Corbin, 1998), such as tracing how routine enactments were evolving over time across projects, or to see how training-learning between pairs of groups different. I further compared labeling routines to "diagnosis routines." Diagnosis routines were clinical interactions between patients and doctors, in which doctors would observe patient signs and symptoms to determine which (if any) disease patients' had. Trainees and senior doctors alike identified diagnosis routines as the source of "patterning" for

labeling routines. While the process coding helped articulate certain aspects of learning in labeling routines that appeared unique (e.g., a "speed dimension"), I noticed during this phase of my analysis that the action-centric approach failed to capture important emotional and cognitive components of learning. For example, several trainees complained that labeling routines were "boring"—a word never used to describe diagnosis routines. Further, the computer-mediated labeling routines seemed to lack an emotional element that several trainees identified as central to their learning process in diagnosis routines. Trainees often shared memorable accounts emotional patient interactions that unfolded in the process of enacting diagnosis routines in the clinic. There were no similar accounts from labeling routines.

Below I use ethnographic descriptions of the retrospective AI development process at Southern to construct a process account of how learning unfolded in labeling routine enactments. I construct a theoretical model which traces how labeling routines led to expert-level learning for AI development (and AI tools), but not for clinical use, in a process that I call "decoupled learning."

FINDINGS

Prior to examining the labeling routine enactments, I briefly cover the process that preceded labeling, in which trainees initially came up with ideas for projects. Trainees perceived retrospective data projects as "safer" than prospective projects, because there were high costs and uncertainties associated with collecting data for prospective AI projects. As one trainee explained to me when talking about coming up with ideas for projects:

> "I wanted to make a tool to predict the Alzheimer's disease (AD) … recently I read a review about this kind of AD [describing a process to detect AD through ocular manifestations]... The paper said a lot of researchers have found that AD patients, they have early signs in the peripheral retina. But if I do this, it's prospective, we collect the data to build this model. It will have a lot of uncertainties… AD patient is not a very huge population … maybe some of them, they memorize something, they already have symptoms, they cannot cooperate with you [to collect data]." *(Interview Transcripts, Trainee 1, Project Amaranth)*

As the trainee here emphasized, sometimes they saw opportunities for projects but ultimately were deemed risky because of the difficulty in collecting data. Several trainees reiterated that

when coming up with an idea, they wanted something that could have real, measurable impact in the clinic:

> "But most important is that you have a good idea. The good idea is one that if you clearly know the requirements of the clinic … And you can, you can ask questions, that is, that are valuable in the clinic. At the same time, it can be solved with the [machine learning] techniques, yeah? … if you ask the engineer to do an impossible task, maybe it will waste his time and your time." *(Interview Transcripts, Trainee 2, Project Malachite)*

As the trainee here illustrates, they judged the value of ideas based on their potential use in clinical settings. Yet at the same time, trainees were responsible for understanding what was technically possible, given the constraints of both data and machine learning algorithms. While trainees were not expected to code (there were software engineers employed fulltime in the AI Department), they were expected to understand the functionality and limitations of AI. In other words, they served as a "bridge" between the technical and medical worlds. Walking into the AI Department's office, the first thing one was greeted with was a bookshelf which represents well this "bridge." I described the shelf early on in my fieldwork:

> *Others are reading books and studying during this time (both AI and ophthalmology). In the primary office is a bookshelf with several books on medicine, AI and various programming languages.* (*AI Office Fieldnotes).*

Trainees then, while not directly responsible for the technical side of AI, were responsible for getting the data prepped for AI. Trainees had to evaluate pre-existing data that was in the hospital and determine what needed to be done to the data in order to prepare it for the machine learning algorithm. Trainees emphasized that this was no easy task:

> "Annotating the data is a dirty job! … If you ever have seen medical data, you will not ask why! It's a mess. Thousands of images that have to be organized [prior to labeling]. Especially with data from electronic health records, they use codes and the information is imperfect, it's not always correct. So, I have to do a standardization process. For example, if you want to know where the patient came from, if your data is from all over the country, some images may have the city name. But some may just have a street name, but no city. So you look up the street name and realize there are thousands of options … First you have to check if something is wrong, and if that thing is serious? Because in big data, something wrong can be acceptable. I try to create a simple principle to extract

some basic information from the data and check, like doing a text analysis." *(Interview Transcripts, Trainee 2, Project Malachite)*

Prior to enacting labeling routines, trainees' biggest task was to assess its viability. This task, referred to as "data prep," took considerable time and effort, as illustrated above. The trainee involved in the above project went on to say that collectively, data prep and labeling constituted "80% of the time" needed to develop AI tools *(Interview Transcripts, Trainee 2, Project Malachite).* Failure to properly prep data before labeling led to major setbacks for projects. Another trainee described how this happened during a project:

> "We make sure that every photo is usable, because … it will like affect the results. So like for example … we found some photos, they have like, dots, we actually don't know where these come from … we found that it was the lens, the lens was dirty … it affects result … the machine recognized these … they think the dirt is a signature of the [specific] disease." *(Interview Transcripts, Trainee 3, Project Lanzones)*

To summarize, prior to enacting labeling routines, trainees had to find project ideas and data that were clinically useful but also (with proper labeling and cleaning) technically viable for machine learning algorithms. As one trainee described to me the process as a whole:

> "AI research requires a lot of data, large amount of data … for AI, you need thousands [of patients]. And then you label the data, clean the data, then train the model. And … validate internally and externally. That is the production line for AI research." *(AI Office Fieldnotes, Trainee 4, Project Olivine)*

The Labeling Routine

Once trainees had prepped their data and had a viable a research idea, their next step was to label data. The labeling routine involved multiple actors engaging in a sequence of interdependent actions, iterated several times in the AI development process, with the end goal of finding a diagnosis. This routine enactment and its patterns are described below.

Recruiting Students Subroutine. As previously mentioned, labeling routines involved multiple actors. This work was considered "dirty," akin to scutwork, and thus it fell onto trainees to manage, rather than senior doctors. As trainees described across several interviews, the workload was very high and required additional assistance:

> "We have to do it [labeling], the senior doctors don't have time. So, I organize teams of students to help. On one project, 5 of us spent 1.5 years on one project, to analyze videos,

there were thousands of them. Each video was 10 frames per second, and 5 minutes per video [3000 frames per video] … in 1.5 years we annotated 1,000 videos." *(Interview Transcripts, Trainee 2, Project Malachite)*

Across all of the projects observed, recruiting students emerged as a subroutine pattern prior to enacting the labeling routine. In this particular project, a trainee recruited four students to assist in labeling. Collectively they spent 18 months labeling three million video frames. Trainees primarily recruited undergraduate medical students to help on the projects. One trainee explained the recruitment process as such:

> "…There is a WeChat group, and in that group many undergraduate students sign up for the group because they are interested in research … there are about two or three hundred members in that WeChat group, and they're all from, as well as you, university… we send a notice like we are carrying on a project and we need like 6 students to help us annotate the images and then like within like 3 minutes, six people [reply]." *(Interview Transcripts, Trainee 5, Project Ecru)*

Depending on the project, trainees would recruit as few as two or as many as 20 students to help with labeling data. Students were paid for their time, and projects could last as little as a few months, or as indicated above, as long as 18 months. While the trainee's description of the WeChat group indicates that there were no shortage of students, the students that were recruited lacked experience in ophthalmology and needed to be trained prior to labeling data.

Training-Learning Pairs in the Labeling Routine. The undergraduates lack of experience was particularly challenging for trainees, as they had a limited amount of time in which they needed to equip students with the skills needed to accurately label data. One trainee recounted their experience onboarding undergraduate students:

> "…They have just finished their basic program. So they have not …[completed] clinical courses like on medicine, surgery and ophthalmology … they are really lacking in technique and knowledge and … the teaching program cannot be finished in a short time, so … we can only tell them 'Look at this, this is the cornea' … then you see this picture, you think this is keratitis … However, keratitis is very different in different patients. It can be white, it can be red, it can be very different. So the annotation is very complex." *(Interview Transcripts, Trainee 6, Project Sienna)*

Students recruited to label data had to be taught how to identify various eye-related disease symptoms specific to the projects that they were assigned to. As indicated above, there was variation in the appearance of certain diseases among different patients, adding to the complexity of teaching the students. Trainees often emphasized that they had to be careful when teaching undergraduate students so that they could readily identify these diseases when they appeared in different forms:

> "For the diabetic retinopathy patients, there may be many exudates from their vessels, they can be seen in the fundus images. They're usually white and may seem like a little bit dirty … the students regard this as like maybe "these are artifacts." … so they see them as bad quality [images], but actually they are not." *(Interview Transcripts, Trainee 5, Project Ecru)*

The challenges in identifying disease were not exclusive to labeling routines, nor were they exclusive to undergraduate students. Take for example from the quote above, a senior doctor explained that retinopathy was very difficult even for clinicians to diagnose:

> "In our field, retinopathy is the most complicated disease. Some are not common or easy to describe. So some doctors can say which part of the retina is abnormal, but don't know specific retinopathy." *(Interview Transcripts, Senior Doctor 1, Project Cordovan)*

Trainees had to commit themselves to learning how to label complex diseases before being able to teach undergraduate students. This was challenging for trainees because they themselves lacked significant clinical experience:

> "Young doctors like me, lacking clinical experience, we read books and ask for help from teachers and colleagues in the clinic. They have more experience than me. The beginning is difficult. But an important point is that even if it is boring, and difficult to train models, after it is complete, the model will help a lot of people and save time. And replace me to some degree." *(AI Office Fieldnotes, Trainee 4, Project Olivine)*

Prior to training undergraduate students, trainees also had to learn. Just as undergraduate students relied on trainees to learn, trainees relied on senior doctors with more clinical experience. This was echoed by several trainees: "Most of my like, ophthalmology knowledge comes from the textbook or like [senior doctor] or lectures." *(Interview Transcripts, Trainee 5, Project Ecru).* Typically, at this stage in their career, trainees would be doing clinical rotations. Instead, their

time in the AI Department counted as a clinical rotation. They saw the opportunity to work in the AI Department as a tradeoff:

> "If I am in the clinic, I know I will see more patients and problems, but then I have no time to do research. But if I am doing research, then it's less time in the clinic. It really is a conflict, you know?" *(Interview Transcripts, Trainee 2, Project Malachite)*

Trainees sacrificed valuable time to learn in the clinic for the opportunity to work on cutting edge AI research. At the same time, trainees felt that labeling routines could also be a source of learning. As the trainee above continued on to say, "I think in the clinic diagnosing a patient is just like how we have to annotate images."

This learning by routine participants at different levels of the professional hierarchy mirrors prior work on legitimate peripheral participation and learning in medicine (e.g., Halsted's method of "see one, do one, teach one" Kotsis and Chung, 2013) in which more senior members train newcomers. Learning in labeling routines differed from previous studies in two particular ways. First, the speed at which trainees had to go from "newcomers" to "old-timers" was much faster with labeling routines. Trainees went from observing and learning how to diagnose (and label) diseases through routine enactments, and on to teaching others in a matter of weeks or months compared to years. Trainees outside the AI Department reported that they would be paired with a more senior classmate in the clinic who would teach them alongside senior doctors. Overtime, those trainees would then become mentors for younger classmates. Yet, for labeling routines, trainees involved were expected to master diagnostic skills quickly so that they could then go on to teach undergraduate students.

Second, labeling routines were a unique medium through which trainees learned. AI models were often "first of their kind," as the trainee above working on *Project Olivine* continued to explain:

> "With AI research, often what you are doing is the first in the world … How to do that research, how to label, research with AI in China and in the world is often lacking a standard for how to do it. So, when we do this research, we have to create a standard, and then we discuss that standard with a lot of [clinical] experts." *(AI Office Fieldnotes, Trainee 4, Project Olivine)*

While trainees relied on senior doctors who had diagnostic expertise in their fields, they also had to make judgement calls about how to translate diagnostic routines into labeling routines

establishing a "gold standard" which they would aim to achieve. In the clinic, it was common for doctors to have variation in the way they diagnose a patient. But AI needs structure, which the trainees had to create prior to performing the labeling routine. Trainees created the gold standard in part through consultation with senior doctors who had clinical experience and diagnostic expertise, but also in part independently through their own understandings of AI development. In an early-stage project which used X-ray images, a trainee described what this meant:

> "They [radiologists] don't have some, very fixed, fixed model of the description, they just write what they want. So this paragraph [pointing to an X-ray report], the description is variable … we can understand this because we can read it, but for AI is a little more difficult … AI always wants structured data." *(Interview Transcripts, Trainee 7, Project Glaucous)*

Senior doctors were able to teach trainees how to perform diagnosis routines, which were the foundation for labeling routines, even though senior doctors did not necessarily understand the "structured elements" of the labeling routine. One senior doctor reflected on their experience teaching trainees involved in AI projects:

> "I notice that they know more about AI technology, but they know little about clinical application or the real world like diagnosis and treatment of diseases. So I think we need to like communicate a lot." *(Interview Transcripts, Senior Doctor 2)*

After consulting with senior doctors, trainees would then learn by labeling a small subset of data. One trainee involved in *Project Malachite* explained to me, "We will first label like 100 or 200 images and test our accuracy to ensure we can label it correctly." I asked the trainee what their accuracy was on these initial images, which they recalled being "93 to 94 percent … maybe some have a higher accuracy, like 97% or 98%, but all of them are above 90 percent, I remember" *(Interview Transcripts, Trainee 8).* For a project to be viable for clinical use, trainees set a threshold above 90% accuracy, "If you actually use this app in clinical settings you should get it to at least more than 90 [percent]" *(Interview Transcripts, Trainee 9, Project Lanzones).* This also meant that training data being input (e.g., data labeled by undergraduate students) had to be at least 90% accurate.

Trainees indicated that even with the steep learning curve of difficult to diagnose diseases, they were able to quickly get to high levels of accuracy in labeling routine performances. They felt that 100 images or "cases" were enough for them to learn most of the

variation that they would see for the entire dataset they were labeling. They even argued that in the case of some rare diseases, 100 images was actually several times more than most clinicians would ever see in a career:

> "Sometimes the dataset is smaller, such as when it is a rare disease. There are fewer cases for a specific disease image. But still, we will see many more than a doctor in a clinic, who might just see a few cases in their life." *(Interview Transcripts, Trainee 2, Project Malachite)*

The above trainee, having just completed a project that included labeling 4,000 images, reflected on their experience:

> "I think the annotation job is very boring. The first 100 you can learn something, but as you continue, there is less and less information you can learn from the images. In Chinese we have a saying, like a joke, 'AI's intelligence depends on the human's work.' [play on the words for "Artificial Intelligence" in Chinese, which can be translated literally as "manual labor intelligence"] Basically, your AI is only as good as the training it gets." *(Interview Transcripts, Trainee 2, Project Malachite)*

Trainees would learn by labeling a small set first, referencing textbooks or asking senior doctors for help on cases they didn't know:

> "At that stage I really learn a lot about ophthalmology... I have to read more textbooks … when I get into the annotation, I read … because [diagnosis in] ophthalmology is really decided on what it looks like, you have to see more and then you know more." *(Interview Transcripts, Trainee 6, Project Sienna)*

Once trainees felt they had a grasp on the topic through labeling a small set of data, they would design an initial training course for undergraduate students. The trainee above, involved in *Project Sienna,* described the course as being very basic:

> "The first time, before the first annotation, we tell them how can you look at an eye in different annotative structures. We have the first course to tell them … which part is which." *(Interview Transcripts, Trainee 6, Project Sienna)*

The trainees focused on teaching undergraduates both technical and medical knowledge in the training session. For example, in a PowerPoint that used to train undergraduates for *Project Ecru,* trainees first presented an overview of the AI development process and how labeling worked from a machine learning perspective. Then they presented specific diagnostic knowledge

(in this case, related to the fundus, or the inner back surface of the eye). They then taught undergraduates how to distinguish between quality images and problematic ones, with several examples such as problems with lighting or blurriness. After training, the undergraduates would be given a small set of data to practice. Undergraduate students accuracy after one round of training was relatively low:

> "For the initial 500 … maybe they can only get, on average maybe 40 percent correct. Yes, by correct I mean all of the items are correct. And or maybe even less than 40 I guess. Yeah, about one-third." *(Interview Transcripts, Trainee 5, Project Ecru)*

Each image would be labeled by two students. The trainee would serve as the "expert." If the two students disagreed on an annotation (i.e., they did not have the same diagnosis), the trainee would evaluate and make the final diagnosis:

> "I analyze their annotations and see if there are any disagreement between the two students. If there are disagreements, then I will check one by one." *(Interview Transcripts, Trainee 5, Project Ecru).*

Trainees had to devote substantial effort to checking the work of undergraduate students:

> "I just, every time, it is like every weekend I receive the package [of labeled data] from the students… that I need to check and to see if there is anything that has to be reannotated. And every time when I begin I will think, 'Ah! My weekend is over.'" *(Interview Transcripts, Trainee 6, Project Sienna)*.

Checking the undergraduate students work was not only essential for reaching the "gold standard," but also for trainees to identify patterns in the mistakes that were being made. Trainees would try to figure out what the recurring issues were, and address them in training:

> "Yes, like I sort out their problems and I made another PowerPoint to teach them. 'OK, in this case, in this case this is what should you do, what should you label,' like that, yeah." *(Interview Transcripts, Trainee 5, Project Ecru)*

By iterating between training sessions, small practice sets and reviewing to catch mistakes, trainees attempted to prevent undergraduate students from learning to misdiagnose in the labeling routine. Another trainee summarized how this iterative process unfolded:

> "Actually, we will offer a small datasets for them to practice … so they can practice and we will check their results. If it's lower accuracy, we will correct their annotations in a different way and teach them how to annotate with higher quality. Then we will, after a

period of annotation, train and correct again, and then we will provide in the final data set for the annotation" *(Interview Transcripts, Trainee 2, Project Malachite)*

In between the training sessions, undergraduate students would still keep in contact with trainees when they were labeling. For example, students would ask for help via WeChat, where they kept group chats with undergraduate students, trainees, and sometimes a senior doctor that was involved in the project. Even though labeling routines did not involve "real-time" interaction between actors, computer-mediated communication was utilized when questions arose. One trainee shared chat logs from their project labeling fundus images, an excerpt of what the communication looked like is below:

Undergraduate student sends unlabeled fundus image to the chat

Undergraduate Student: *[Trainee], sorry, how should I label this image?*

Senior Doctor [first to reply about one hour later]: *This is a lesion known as central retinal vein occlusion, CRVO*

Undergraduate Student: *Thank you! (AI Clinic Archival Documents)*

Trainees across several projects developed consensus that the undergraduate students could achieve approximately 80% accuracy, and that to achieve that level, they needed to have at least three rounds of training:

"I think if you want to get more accurate results, maybe you have to offer three or more rounds of training" *(Interview Transcripts, Trainee 2, Project Malachite)*

"Every picture has to be annotated by two [students]… So, the first time about 50 percent need to be reannotated [by the trainees], and the second and third time about 20 percent … What I think is that when we train them, if we really want to train them to an accuracy of about 90 percent or 95 percent, maybe by then our work has already finished …" *(Interview Transcripts, Trainee 6, Project Sienna)*

"Did he [trainee] tell you about the training sessions? Actually for 2 weekends we had the students come here and trained them. Their accuracy is about 80%. They are not bad." (*Interview Transcripts, Senior Doctor 1, Project Ecru)*

Once undergraduate students had reached this limit (in accuracy and number of training sessions), it was up to the trainees to correct the rest of the images and to get the accuracy above

the 90% threshold that they had set. The following is an excerpt of fieldnotes that demonstrates how the labeling routine was enacted by trainees:

10:00am

The trainee opens a patient's medical record on his laptop [I can see this is patient number 1218—he has labeled data for 1218 eyes so far]. He enters the patient's name into an Excel spreadsheet along with their date of birth [he explains that later they anonymize the patient data in the spreadsheet]. He checks "0" under the diagnosis column [0 indicating is no diagnosis for the clinic]. The trainee creates a new file folder using the patient's anonymous identification number, and saves the patient images in the folder. He scrolls through all the images rapidly; he is scrolling so quickly that it almost looks like an animation. He then fills out a row of labels in Excel [indicating clinical manifestations and diagnosis], and moves to the next eye.

I ask him, "Do you ever get the same patient twice?" He continues to input values for the second eye as he replies to me, "Sometimes, it should all be in one folder. Here, if they come back [pointing to the screen] then there will be another tag with the date. But sometimes they just create a new folder."

The trainee opens up a new patient's folder. He looks through the photos and scrolls quickly again (Eye #1220). He opens Excel again and selects a number from a dropdown list of possible diagnoses that he has created. He goes back to the images. I notice he is labeling very quickly today. He scrolls through the images, and then goes back to Excel and records information. Next he copy and pastes the patient's name and basic information into the next row for the patient's second eye.

10:13am

The trainee opens the next set of images for the patients other eye. He leans closer to the computer, scrolls back and forth through the images, before going to Excel and recording the results.

[I sketch in my fieldnotes how his computer screen is setup: images of the patient's eyes are set up next to one another with the "tags" (labels) written in English above]

The trainee silently repeats back a string of numbers—the patient's anonymous identification number—as he types the number in as the name of a new folder. He then exports the patient's photos in the folder. He opens the images and taps through … the trainee then opens Excel, pauses, and enters in a series of values …

He sits back and says to me, "When we train the model, the AI treats one eye as one data, not linking the two together. I think that could be a problem. Because if you have a disease in one eye, the probability you will have it in the other eye is higher"

…

10:32am

The trainee is scrolling through another image (Eye #1228) [someone has attempted to label this one]. He pauses on one picture and zooms in. I ask what he is looking for. He tells me, "I want to figure out what is the white spot here [pointing to a white spot on the image], it could be a drusen [lipid deposit on the retina] … he confirms to himself that he thinks it is, and inputs the label.

(AI Office Fieldnotes, Trainee 7, Project Icterine)

As demonstrated in the above excerpt, the labeling routine was a computer-mediated process in which the trainee would evaluate patient medical records and images to come up with a diagnosis. The trainee was able to move very quickly from "patient" to "patient" (or more precisely, from "eye" to "eye"). There was no downtime in between patients, as it was a repetitive pattern of actions in which the trainee would consolidate patient information, assign an anonymous number, and then evaluate the images. For easy cases, trainees were able to rapidly label images (as demonstrated above). More difficult cases (such as the last example from the excerpt) took more time and effort for trainees.

Struggles in the Clinic

Many trainees perceived labeling routines and diagnosis routines which were performed in clinical work to be very similar. Trainees commented on how they felt they learned a lot in

labeling routines by being exposed to different manifestations of disease: "You actually can see a lot of features that are very different from the textbook [when labeling]" *(Interview Transcripts, Trainee 6, Project Sienna)*. Despite achieving high levels of accuracy in labeling routine performances, trainees began to experience unanticipated challenges as they spent more time in clinical rotations. Several trainees spoke about the emotional toll of patient interactions that they were not exposed to when enacting the labeling routine. One trainee recalled when they had to do a rotation in a department that dealt with suspected tumor cases:

> "Do you know this disease? [cancer that affects children] It is a very cruel disease … if the patients have this disease, you have to remove their eye … it is difficult to contact the parents and tell them the truth … I think that is a very tough task for me … you have to tell them about the condition of their children and then of course for many parents it is very shocking … These children are good, just like angels … I can't imagine, when I first go to the clinic, I cannot imagine what the parents or the children would be like in the future … and when I finish surgery … I'm afraid to go back to the clinic to see the patient." *(Interview Transcripts, Trainee 6, Project Sienna)*

While the goal of labeling routines was to find a diagnosis, the trainees never had to inform patients of their diagnosis, as was the case in clinical work when enacting diagnosis routines in the clinic. Senior doctors also noticed that trainees were caught off guard by this need to communicate with patients:

> "Their clinical experiences are not the same [between trainees in the AI department and others] … on the clinical floors you also need to communicate with the patient and tell them how their surgery is … or what they need to pay attention to after the surgery. The students here, they don't have enough clinical experience, so in this case, sometimes they don't know how to explain to the patient…" *(Interview Transcripts, Senior Doctor 2)*

Labeling routines were computer-mediated and largely cognitive activities—trainees (and students) would look at images and medicals records on their computer screen and assign a diagnosis based on what they observed. This diagnosis was recorded electronically and not communicated with patients, unlike diagnosis routines which involved patient communication.

Labeling routines also prioritized speed over accuracy. While 90% accuracy was good for an AI model, a 10% possibility for error still represented a major risk for doctors. As another senior doctor explained to me:

> "In another field if you are wrong in 1% of cases, it's not a big deal. But in medicine 1% is very costly, so a human doctor has to offer a guarantee when they give a diagnosis." *(Interview Transcripts, Senior Doctor 1, Project Cordovan)*

One trainee, having just finished a 3-month clinical rotation, reflected on their experience enacting labeling routines, "I think when you are doing AI projects, you have learned to make mistakes" *(Interview Transcripts, Trainee 8, Project Malachite)*. Another trainee who had never worked in the AI Department or performed labeling routines, described how they approached their clinical work at the beginning:

> "I was very nervous, I don't want to miss any symptoms, so I was very slow. But later on I just got used to it. Yeah, it was better. So right now I'm more efficient and quicker." *(Interview Transcripts, Trainee 10)*

I asked for a specific example of what their experience was like at the beginning, and they explained how being slow was actually helpful as they were learning:

> "Once there was this patient that came here many times … they just don't know what the problem was, and this time, when I was looking at her, and I finally found out the reason for her disease … I was still slow then and examined everything … I looked very carefully and then I found like 'sand' in her cornea. But it was not very obvious, I only saw it because I did staining … if you are not in a rush so you can find out more …" *(Interview Transcripts, Trainee 10)*

In comparison with trainees that performed labeling routines, trainees who performed diagnosis routines were focused on accuracy over speed. Trainees who had performed labeling routines prior to extensive clinic work struggled with diagnosing routines that took more time and involved patient interaction.

Communication was a core part of the diagnosis routine enactment in the clinic. In addition to the above examples which focus on the emotional elements of comforting patients and their family, communication and patient interaction was an essential part of reaching an accurate diagnosis in the clinic. As one trainee explained in comparing labeling routines with diagnosis routines:

> "Somethings you can just judge by an image, like a hemorrhage. You see, the red is very obvious, it's very easy or something like a retinal detachment. Like these kinds of diseases, you can judge just by an image. But like glaucoma, you can't judge it—like a

> diagnosis of glaucoma just by images, I can only label as suspected glaucoma [in AI] … if I suspected these people may have glaucoma [in the clinic], I will give them a call, ask them back and check their intraocular pressure." *(Interview Transcripts, Trainee 1, Project Amaranth)*

Trainees picked up on these contrasts between labeling and diagnosing routines (in the clinic) after experiencing both. In the clinic, if a doctor suspected a problem but did not have enough information, they could run additional tests, or ask for more information. But in the labeling routine, trainees had to make inferences based on the data presented to them. They began to suspect that this limitation may be the reason that the AI tools being developed were not performing well in the clinic:

> "Because in the clinic, if you want some more data, we can ask patient to do to provide it. But with AI research the data is just there, they can't be expanded." *(Interview Transcripts, Trainee 7, Project Icterine)*

Another trainee explained the implications of this limitation:

> "So what I want to say is that, even some doctors from the AI Clinic, they will be suspicious … because sometimes, we don't have the ground truth about the symptoms. So some of the research work looks very good, but in the real world it does not work … When we train the system, we will think about 'How can this be the ground truth?' Sometimes we have a lot of descriptions, examinations but we cannot get the real diagnosis [when labeling]." *(Interview Transcripts, Trainee 6, Project Sienna)*

I asked the trainee why they could not get the real diagnosis and she attributed it to the way data was collected in the clinic:

> "Because many, many patients come to our clinic. Examinations cost a lot of money. If the patient does not decide to get treatment, then the doctors will not do too much to examine them … most of the patients, they do not get a definitive diagnosis at all, for their lifetime … Sometimes doctors just use medicine, and then their visual acuity gets better and then the patient is like 'Oh that's good!' They don't care about what their disease is. So that is the question. We do not have the real diagnosis for each patient … So the AI system cannot have the true diagnosis … It will limit the use in the clinic, in the real world." *(Interview Transcripts, Trainee 6, Project Sienna)*

The trainee explained that doctors often stopped short of collecting enough information to make a definitive diagnosis in the clinic. Instead, they relied on guesswork, seeing if medicine would lead to improvements. If the patients symptoms improved, they would not try to run anymore tests. As a result, the data that trainees had to work with was limited. Often, there was no definitive diagnosis from the clinic, which in turn limited the ability of AI tools they developed to provide a "true diagnosis."

DISCUSSION

A Process Model Decoupled Learning in Labeling Routine Performances

Below I describe a process model to illustrate how trainees were able to achieve expert-levels of accuracy in labeling routines, but struggled to reach these same levels of accuracy in diagnosis routines with real patients in the clinic. Trainees and senior doctors were able to see many parallels between diagnosis routines and labeling routines. In fact, senior doctors, who had never performed labeling routines, were able to teach trainees the necessary skills to perform labeling routines based on their diagnostic expertise. Diagnosis routines, at a minimum, required a doctor, patient, and data as artifacts (images, medical records, etc.). In the diagnosis routine performance, doctors would interact with patients, collect data, and use that data to come up with a diagnosis. Labeling routine performances removed patients from the equation, only their data remained. The labeling routine performance was computer-mediated and involved trainees (or students) who would use available patient data to come up with a diagnosis. Despite the surface-level similarities, and the ability of doctors to transfer their expertise, I argue that there are two main reasons that the repetitive labeling routine performances led to "decoupled learning," and made it difficult for trainees to enact diagnosis routines in the clinic.

First, "speed" was a recurring theme in labeling routines, which had several implications for the way that trainees learned. Trainees had to learn to diagnosis quickly, because they were also responsible for training undergraduate students. While the initial patterning of the diagnosis routines taught by senior doctors were grounded in the "real-world" clinic, trainees performance was in computer-mediated labeling routines. Typically, trainees would also "learn by doing" in the clinic, enacting diagnosis routines alongside doctors (or more advanced trainees). Yet for AI development, trainees instead learned through computer-mediated labeling routines. This removed the physical, performative aspect of the diagnosis routine that involved multiple actors. As trainees repeated the labeling routine performance and increased their accuracy, their

patterning of this individual, computer-mediated cognitive routine enactment was reenforced. This drew trainees further away from the initial patterning presented by senior doctors, while the increasing accuracy reenforced their belief that they were achieving expertise. Speed also emerged as a theme in labeling routines in the sense that labeling routines prioritized speed over accuracy. Given that trainees had to annotate thousands of images, they were willing to concede that errors would happen. As noted above, doctors in the clinic, particularly trainees, were careful not to make mistakes, because they were dealing with patients' lives. This willingness to accept a certain margin for mistakes continued to drive a wedge between the mutual constitution of performing and patterning of labeling and diagnosis routines.

Second, the theme of "real-time" interaction emerged as important in understanding the decoupled learning in labeling routines. Diagnosis routine performances featured multiple actors in "real-time" (e.g., at least one doctor and one patient) in a clinic. Labeling routines featured staggered interaction with other actors. As noted in the findings, trainees (and undergraduate students) did interact, primarily in training sessions and via WeChat with questions. While the training sessions were "real-time" interactions, they were intermittent, and the labeling routines performed in between sessions were largely individual and cognitive (one trainee with data on the computer). When trainees were placed into the clinic, they struggled with "real-time" multiple actor components of the diagnosis routine enactment that they had not faced when labeling, such as how to communicate with patients and deal with the emotional toll that diagnosis routines regularly produce.

Contributions and Implications

In studying how labeling routines influenced the way new entrants into a profession learn, my study offers several contributions to scholarship on routine dynamics and situated learning. Carnegie School scholars (Cyert & March, 1963; March & Simon, 1958; Simon, 1947; Nelson & Winter, 1982) primarily saw routines as cognitive patterns that could "stabilize" an organization (Salvato & Rerup, 2011). Routine dynamics, by introducing a performative aspect, considered routines to be a source of stability and change in organizations (Feldman & Pentland, 2003). Rather than viewing stability and change as antithetical, routine dynamics scholars have demonstrated that they are mutually constituted (Feldman et al. 2021: 3). As Feldman (2016) argues, this is closely linked to the notion that patterning and performing aspects of routines are also mutually constituted. Routine dynamics scholarship has backgrounded the cognitive aspects

of routines (Lazaric, 2021; Levinthal & Rerup, 2006; Rerup & Spencer, 2021), instead emphasizing the performative "actions" of routines (Pentland & Goh, 2019).

I found that over the course of thousands of repetitions of computer-mediated labeling routine enactments, routine performances largely became cognitive in nature. While labeling routine performances appeared to stabilize, as Carnegie Scholars would predict, this surface level stabilization was masking a "decoupling" underneath. Each iteration of the labeling routine drifted further away from the physical and emotional patterning of clinical diagnosis routines. The emphasis on speed in labeling routine performances disrupted the mutual constitution of patterning and performing of diagnosis routines for trainees in the clinic. This account challenges the idea that learning is based on the outcome of routine performances involving multiple actors (Feldman & Pentland, 2003; Miller, Pentland, & Choi, 2012), as learning in this case features instances that are largely cognitive and individual. As a result, this excessive cognitive patterning produced rigid routines that were less malleable even when trainees returned to clinical settings.

In studies of situated learning, the transition from legitimate peripheral participation to full participation is "never unproblematic" (Lave & Wenger, 1991: 116). Adopting a routine dynamics view of learning as a process of patterning (Rerup & Spencer, 2021) helps resolve the "expertise paradox" that I observed in my site, in which trainees, despite being experts at labeling diseases, were unable to diagnose the same diseases when in the clinic. Not only can virtual simulations increase cognition representations of learning, but I demonstrate that after thousands of repetitions, they can impede on the social aspects of situated learning. Much like AI models can be trained excessively, leading to "overfitting," trainees engaging in extensive repetitions of labeling routines struggled to generalize skills when enacting diagnosis routines.

Conclusion

In capturing the decoupled learning process from labeling routines, I highlight why scholars should focus on the development of AI to move theory forward (Bailey & Barley, 2020). An important implication of this study is that AI development can lead to the deskilling of humans. Yet, a perhaps counterintuitive lesson of this study is that AI development *can* be beneficial to facilitating learning for new entrants to a profession—in moderation. If more trainees had been involved in labeling routine performances, they may have been able to derive more benefits of learning without reaching the detrimental levels of repeated labeling routine performances.

REFERENCES

Bailey, D. E., & Barley, S. R. (2011). Teaching-learning ecologies: Mapping the environment to structure through action. *Organization Science*, *22*(1), 262-285.

Bailey, D. E., & Barley, S. R. (2020). Beyond design and use: How scholars should study intelligent technologies. *Information and Organization*, *30*(2), 100286.

Bailey, D. E., Leonardi, P. M., & Barley, S. R. (2012). The lure of the virtual. *Organization Science*, *23*(5), 1485-1504.

Barley, S. R. & Beane, M. (2020). How Should We Study Intelligent Technologies' Implications for Work and Employment? In S. R. Barley *Work and Technological Change* (pp. 69–115). Oxford University Press, USA.

Barley, S. R., & Kunda, G. (2001). Bringing work back in. *Organization Science*, *12*(1), 76-95.

Beam, A. L., & Kohane, I. S. (2016). Translating artificial intelligence into clinical care. *Jama*, *316*(22), 2368-2369.

Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, *64*(1), 87-123.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020, April). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-12).

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization Science*, *2*(1), 40-57.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534.

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, *49*(5), 897-918.

Cohen, M. D., & Bacdayan, P. (1994). Organizational routines are stored as procedural memory: Evidence from a laboratory study. *Organization Science*, *5*(4), 554-568.

Cook, S. D., & Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization science*, *10*(4), 381-400.

March, J. G. (1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ: Prentice-Hall.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE.

Edmondson, A. C., Bohmer, R. M., & Pisano, G. P. (2001). Disrupted routines: Team learning and new technology implementation in hospitals. *Administrative Science Quarterly*, *46*(4), 685-716.

Eisenhardt, K. M., & Tabrizi, B. N. (1995). Accelerating adaptive processes: Product innovation in the global computer industry. *Administrative Science Quarterly*, 84-110.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, *28*(1), 62-70.

Feldman, M. (2016). Routines as Process: Past, Present, and Future. In In J. Howard-Grenville, C. Rerup, A. Langley, and H. Tsoukas (eds.), *Organizational Routines: How They Are Created, Maintained, and Changed* (pp. 23-46). Oxford University Press.

Feldman, M. S., & Pentland, B. T. (2003). Reconceptualizing organizational routines as a source of flexibility and change. *Administrative Science Quarterly*, *48*(1), 94-118.
Feldman, M. S., Pentland, B. T., D'Adderio, L., Dittrich, K., Rerup, C., & Seidl, D. (2021). What is Routine Dynamics? In M.S. Feldman, B.T. Pentland, L. D'Adderio, K. Dittrich K., C. Rerup , and D. Seidl (eds.). *Cambridge Handbook of Routine Dynamics* (pp. 1-18). Cambridge: Cambridge University Press.
Goh, K. T., & Pentland, B. T. (2019). From actions to paths to patterning: Toward a dynamic theory of patterning in routines. *Academy of Management Journal*, *62*(6), 1901-1929.
Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. HMH Books.
Hinton, G. (2016). "On Radiology." Comments presented at the 2016 Machine Learning and Market for Intelligence Conference, Toronto, ON. https://www.youtube.com/watch?v=2HMPRXstSvQ
Hirschauer, S. (1991). The manufacture of bodies in surgery. *Social Studies of Science*, *21*(2), 279-319.
Huising, R. (2015). To hive or to hold? Producing professional authority through scut work. *Administrative Science Quarterly*, *60*(2), 263-299.
Johnson, E. (2007). Surgical simulators and simulated surgeons: Reconstituting medical practice and practitioners in simulations. *Social Studies of Science*, *37*(4), 585-608.
Kellogg, K. C., Myers, J. E., Gainer, L., & Singer, S. J. (2021). Moving violations: Pairing an illegitimate learning hierarchy with trainee status mobility for acquiring new skills when traditional expertise erodes. *Organization Science*, *32*(1), 181-209.
Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366-410.
Koschmann, T., LeBaron, C., Goodwin, C., & Feltovich, P. (2011). "Can you see the cystic artery yet?" A simple matter of trust. *Journal of Pragmatics*, *43*(2), 521-541.
Kotsis, S. V., & Chung, K. C. (2013). Application of see one, do one, teach one concept in surgical training. *Plastic and reconstructive surgery*, *131*(5), 1194.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge university press.
Lazaric, N. (2008). Routines and routinization: an exploration of some micro-cognitive foundations. In M. C. Becker (Ed.), *Handbook of Organizational Routines* (pp. 205-227). Cheltenham: Edward Elgar.
Lazaric, N. (2021). Cognition and Routine Dynamics. In M.S. Feldman, B.T. Pentland, L. D'Adderio, K. Dittrich K., C. Rerup , and D. Seidl (eds.). *Cambridge Handbook of Routine Dynamics* (pp. 255-265). Cambridge: Cambridge University Press.
Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *Management Information Systems Quarterly*, *45*(3), 1501-1526.
Levinthal, D., & Rerup, C. (2006). Crossing an apparent chasm: Bridging mindful and less-mindful perspectives on organizational learning. *Organization Science*, *17*(4), 502-513.
March, J. G., & Simon, H. A. (1958). Organizations. John Wiley & Sons, New York.
Metz, C. (2019). A.I. is learning from humans. Many humans. *The New York Times*. https://www.nytimes.com/2019/08/16/technology/ai-humans.html
Miller, K. D., Pentland, B. T., & Choi, S. (2012). Dynamics of performing and remembering organizational routines. *Journal of Management Studies*, *49*(8), 1536-1558.

Nelson, R. R., & Winter, S. G. Winter (1982). *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press.

Parmigiani, A., & Howard-Grenville, J. (2011). Routines revisited: Exploring the capabilities and practice perspectives. *Academy of Management annals*, *5*(1), 413-453.

Pine, K. H., & Bossen, C. (2020). Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society*, *7*(2), 2053951720965616.

Pine, Kathleen H., et al. "Innovations in clinical documentation integrity practice: Continual adaptation in a data-intensive healthcare organisation." *Health Information Management Journal* (2021): 18333583211067845.

Prentice, R. (2005). The anatomy of a surgical simulation: The mutual articulation of bodies in and through the machine. *Social Studies of Science*, *35*(6), 837-866.

Rerup, C. & Spencer, B. (2021). Carnegie School Experiential Learning and Routine Dynamics. In M.S. Feldman, B.T. Pentland, L. D'Adderio, K. Dittrich K., C. Rerup , and D. Seidl (eds.). *Cambridge Handbook of Routine Dynamics* (pp. 445–459). Cambridge: Cambridge University Press.

Sachs, S. E. (2020). The algorithm at work? Explanation and repair in the enactment of similarity in art data. *Information, Communication & Society*, *23*(11), 1689-1705.

Saldaña, J. (2014). Coding and analysis strategies. In P. Leavy (ed.) *The Oxford Handbook of Qualitative Research* (pp. 581–605). Oxford: Oxford University Press.

Salvato, C., & Rerup, C. (2011). Beyond collective entities: Multilevel research on organizational routines and capabilities. *Journal of Management*, *37*(2), 468-490.

Simon, H. A. (1947). *Administrative Behavior*. Cambridge: Cambridge University Press

Strauss, A. (1978). A social world perspective. In N. Denzin (ed.), *Studies in Symbolic Interaction* (pp. 119-128). Greenwich, CT: JAI Press.

Strauss, A., & Corbin, J. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 2d ed. Thousand Oaks, CA: Sage.

Topol, E. J. (2020). Welcoming new guidelines for AI clinical research. *Nature medicine*, *26*(9), 1318-1320.

Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge, UK: Cambridge University Press.