Mucha, Tomasz; Seppälä, Jane; Puraskivi, Henrik

Context Changes and the Performance of a Learning Human-in-the-loop System: A Case Study of Automatic Speech Recognition Use in Medical Transcription

Published in:
Proceedings of the 56th Hawaii International Conference on System Sciences

Published: 01/01/2023

Document Version
Publisher's PDF, also known as Version of record

Context Changes and the Performance of a Learning Human-in-the-loop System: A Case Study of Automatic Speech Recognition Use in Medical Transcription

Tomasz Mucha
Aalto University School of Science
tomasz.mucha@aalto.fi

Jane Seppälä
Aalto University School of Science
jane.seppala@aalto.fi

Henrik Puraskivi
Inscripta Oy
henrik.puraskivi@inscripta.io

Abstract

The paper presents how organizational practices enable the improvement and maintenance of task performance in a learning human-in-the-loop system exposed to a wide range of context changes. We investigate how the case company tripled the efficiency of medical transcribers by leveraging its machine learning-based automatic speech recognition technology. We find that the focal system operated across stable, drifting, and jumping contexts. Despite changes, it continued to improve or maintained performance thanks to two sets of organizational practices aligning it with the context: extending and refining. This paper makes two key contributions: It shows the importance of considering context changes in the design and operation of learning human-in-the-loop systems. Our empirical findings help with resolving some contradictory outcomes of the recent conceptual work. Secondly, we show that context alignment practices are situated at the sociotechnical system level and, thus, are not just technical solution nor can be detached from social elements.

Keywords: human-in-the-loop, machine learning, artificial intelligence, task performance.

1. Introduction

Organizations across the world are keen to improve their performance in executing various tasks by leveraging machine learning (ML) technologies (Ransbotham et al., 2020). ML-based solutions are frequently deployed as human-in-the-loop systems (henceforth systems or loops), where human agents monitor the automatically generated outputs and, if needed, intervene (Benbya et al., 2021). Records of these interventions serve as training data in subsequent ML model training rounds (Paterson et al., 2021). Thus, the systems are learning from the accumulated experience. This learning capability renders the potential for creating a virtuous cycle allowing ML to continually improve, hence contributing to increased performance of the entire loop (Paterson et al., 2021; Ransbotham et al., 2020). This notion, however, rests on the assumption that task context remains stable, so that past training data is representative of the future (Benbya et al., 2021). With the recent examples of major disruptions, such as COVID-19, and mounting evidence about ML performance deterioration due to drift in the data (Kim et al., 2022), organizations relying on learning human-in-the-loop systems cannot ignore the challenges posed by context changes.

Against this backdrop, we ask: *"How learning human-in-the-loop systems improve and maintain performance in a changing task context?"* This critical question, to our understanding, has not been comprehensively addressed thus far. Insights from conceptual and simulation-based research are contradictory. Balasubramanian and colleagues (2020) argue that ML should substitute human decision-making in case of major context changes, while Sturm and colleagues (2021) suggest that high-level of human involvement is needed in turbulent environments. Apart from bringing empirical evidence to move the discussion forward, we uncover nuance of organizational practices, thus answering the "how" part of the question.

Our contribution to this emerging discussion builds on a qualitative case study set in the context of medical transcriptions – conversion of audio files with patient notes recorded by doctors into text. We investigate the performance changes and organizational practices of a learning human-in-the-loop system operated by a company offering transcription services to medical clinics. Our study provides unique insights, because the focal system has tripled performance and has been operating continuously for 4 years across a wide range of context changes, such as new doctors, clinics, medical fields, languages, and sometimes dealing with audio recordings in other sectors than healthcare. Our insights are equally relevant to practitioners who need to align their human-in-the-loop systems with context.

URI: https://hdl.handle.net/10125/103014
978-0-9981331-6-4

2. Background

Research on learning human-in-the-loop systems is in its infancy. IS and organizational scholars have been studying questions pertaining to such systems or related phenomena predominantly in the research streams dealing with automation and augmentation (Raisch & Krakowski, 2021; Strich et al., 2021; Teodorescu et al., 2021), managing artificial intelligence (Berente et al., 2021; Grønsund & Aanestad, 2020; Lyytinen et al., 2020; Mucha et al., 2022), IS impact on task performance (Sturm & Peters, 2020), organizational learning (Sturm, Gerlach, et al., 2021; Sturm, Koppe, et al., 2021), and human-computer interaction (Budd et al., 2021; Fügener et al., 2021; Ge et al., 2021; Kim et al., 2022). While a comprehensive literature review is beyond the scope of this paper, we provide here an overview of selected work which allows us to position our study within the recent body of research on organizational use of human-in-the-loop systems and ML, as well as their impact on task performance.

A growing share of organizational tasks could be automated (Parasuraman et al., 2000). Yet, partial automation leading to the creation of a human-in-the-loop system is nevertheless very common, because of cost, performance, legal, ethical, safety, and other constraints (Fügener et al., 2021). These systems do not simply optimize the preceding activities and workflows by (partially) automating some of the constituent tasks within the existing workflows, but rather substitute and/or change them in often unintended and unanticipated ways (Parasuraman et al., 2000). This is particularly the case with systems based on ML technologies, because of their heightened levels of agency, ability to learn, inscrutability, and capacity to form new human-machine configurations (Benbya et al., 2021; Berente et al., 2021). Hence, learning human-in-the-loop systems present a novel area of research, which is of high practical relevance. Such systems are currently in evaluation or production stage in domains as diverse as stock market trading (Sturm, Koppe, et al., 2021), selecting goods for customs inspection (Kim et al., 2022), evaluating job applicants (van den Broek et al., 2021), manufacturing quality inspection (Koppe & Schatz, 2021), consumer lending (Strich et al., 2021), or analysis of maritime commodity shipments (Grønsund & Aanestad, 2020).

In the remainder of this section, we highlight four pivotal aspects of these systems and their impact that require careful attention, if we are to accumulate knowledge and understanding in this nascent strand of literature – learning, role of humans, task performance, and task context. First, while learning is one of the key defining characteristics of these systems (Berente et al., 2021; Lyytinen et al., 2020), it is important to draw a distinction between systems which learn in the development phase only and those which continue to learn and change throughout their operational life. ML technology can be at the core of both types of systems (Paterson et al., 2021). The inherent dynamism of the latter renders them meaningfully different and most interesting. Accordingly, practitioners (Ransbotham et al., 2020) and conceptual work (Balasubramanian et al., 2020; Lyytinen et al., 2020; Raisch & Krakowski, 2021) have concentrated on continually learning systems. However, in-depth empirical case studies of such systems are still very scarce (notable examples are: Grønsund & Aanestad, 2020; Sturm, Koppe, et al., 2021). Second, human actors can play different roles in various configurations of human-in-the-loop systems (Grønsund & Aanestad, 2020). In some systems humans decide on whether to use the outputs from ML technology (Fügener et al., 2021; Ge et al., 2021) or are expected/forced to follow them (Strich et al., 2021). Practitioners highlight that systems where humans provide feedback and teach the technology generate more value over time (Ransbotham et al., 2020). Third, in line with the long-standing interest in the impact of IS on task performance (Goodhue & Thompson, 1995; Sun et al., 2019), there is an emerging research on efficiency, effectiveness, and knowledge accumulation within these systems (Fügener et al., 2021; Ge et al., 2021; Sturm & Peters, 2020). While many find positive performance impacts (Grønsund & Aanestad, 2020; Strich et al., 2021; Sturm, Koppe, et al., 2021), this is not universally true (Fügener et al., 2021; Ge et al., 2021; van den Broek et al., 2020). There is still very little research on how these impacts are generated in real organizational setting and over extended periods of time. Empirical studies often cover only the initial system development stage (Grønsund & Aanestad, 2020; Koppe & Schatz, 2021), short timeframe (Sturm, Koppe, et al., 2021), or lab experiments (Fügener et al., 2021). Forth, task context is often overlooked, despite being an important characteristic driving human-machine collaboration (Baird & Maruping, 2021; Mucha et al., 2022) and team performance (Ilgen et al., 2005). Even though ML algorithm performance routinely degrades with task context changes (Kim et al., 2022) or minor variations in task workflow (Budd et al., 2021), most research on learning human-in-the-loop systems are conducted in static contexts (Fügener et al., 2021) or periods of moderate change only (Sturm, Koppe, et al., 2021). These limitations of present work inhibit our understanding of how organizations successfully deal with context changes while protecting or improving the performance of learning human-in-the-loop systems. The study we present next, addresses these shortcomings.

3. Research setting and methodology

To study how continually learning human-in-the-loop systems operating in a real organizational setting improve and maintain performance in changing task context we rely on inductive single case study methodology (Yin, 2009). This approach allows us to study complex processes with rich and nuanced insights (Graebner et al., 2012) and generate answers to "how" and "why" questions (Eisenhardt, 1989). Furthermore, qualitative case studies are suitable for approaching novel phenomena, which are multidisciplinary, information rich and cross multiple levels of analysis (Graebner et al., 2012; Yin, 2009).

In this study we rely on purposive sampling to select our case organization – Inscripta[1]. The company presents multiple characteristics that make it an ideal candidate for gaining insight needed to address our research question. Inscripta is a start-up company with a sole focus of developing and commercializing its ML-based transcription service (automatic speech recognition or ASR). This focus is beneficial for our research, because virtually all Inscripta's organizational practices and attention revolve around the subject of our study. This means that all interview respondents are deeply involved in activities related to ASR and, at the same time, there are no other issues or business activities that would unnecessarily complexify or obscure our access to relevant data. Next, Inscripta's ASR is continually learning through feedback from transcribers, who audit and, if needed, modify automatic output. Thus, our case is representative of modern practices prevailing in the industry (Ransbotham et al., 2020). Also, this case setting provides an excellent opportunity to study the impact of changes in task context on task performance. On the one hand, the task for which the focal system has been developed presents a high level of uniformity along some dimensions. Thus far, the company has been primarily working with medical service providers – generating patient notes based on audio files recorded by doctors. Thanks to bulk of Inscripta's past work constituting medical transcriptions the organizational practices and the resulting task performance are comparable across time. On the other hand, the case company has been actively and successfully expanding, thus covering an increasing range of doctors, clinics, medical fields, and sometimes dealing with audio recordings in other sectors than healthcare. Such context changes render most ML algorithms used in medical domain unreliable and underperforming (Budd et al., 2021). Thus, this case illustrates performance impact of context changes and the organizational practices developed to deal with these. Finally, Inscripta has been delivering medical transcriptions starting from 2017 and they have utilized for this task internally developed system since 2018. This means that our study offers a unique insight into organizational practices over a long period of time and across various types of context changes. Thus, our study extends past research on learning human-in-the-loop systems in several important ways.

Data collection took place in the spring of 2022. We conducted 10 semi-structured interviews (referred to as I1-I10), out of which 5 were conducted with doctors and transcribers not involved with the case company to build a broader understanding of the medical transcription domain and 5 were conducted with Inscripta employees: management, transcribers, and operative personnel. The interviews lasted between 25 and 100 minutes. The theme of the interviews was the impact of ASR technology on the medical transcription work. The interviews were transcribed using speech-to-text technology and supported by manual transcription where needed. We also collected 38 public documents including LinkedIn posts and publications by the company as well as news articles mentioning the company (referred to as Doc1-Doc38). When combined into a single document the full text of transcripts and archival materials extended to over 103,000 words (over 300 pages, 12-point Times New Roman, single-spaced).

We started data analysis while still gathering data to be able to guide our focus and update the interview questions. Initially we utilized open coding for all the materials. This coding was done independently and then jointly by first two authors utilizing Atlas.ti Web qualitative analysis software. We ended up with over 500 codes that represented the first-order concepts, such as "4x performance improvement", "pathology", and "quality is important". We then grouped these into 14 second-order themes by comparing and contrasting the first-order themes and the quotes they represented. These included themes such as "context switching", "algorithm learning", and "human learning". We recognized that "context switching" represented an important theme that heavily influenced task performance. We started mapping activities and practices that seemed relevant to this theme. We also mapped the differences in workflows before and after ASR introduction, and analyzed changes in the process. At this stage, we started testing different models as guided by literature on task performance (Goodhue & Thompson, 1995), organizational learning (Sturm, Gerlach, et al., 2021), and sociotechnical systems (Winter et al., 2014). We ultimately arrived at the findings presented below.

[1] https://inscripta.io/

4. Findings

4.1. Transcription task performance

Since its founding in 2017, Inscripta has been offering medical transcriptions. Initially, the transcripts were generated manually by transcribers who listened to audio recordings and typed the text, as presented in Figure 1A. Starting from 2018, the company has rapidly transitioned to utilizing ASR with human transcribers-in-the-loop as the primary mode of transcription task execution (Figure 1B). This transition resulted in approximately tripling the efficiency of transcribers while maintaining or, potentially, improving the quality of transcripts.

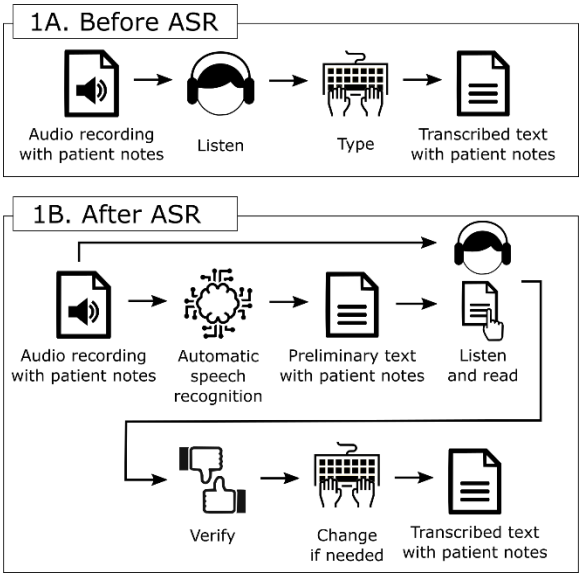

Figure 1. Transcription process before (1A) and after (1B) introducing ASR.

The quality or accuracy of medical transcription is one of the indicators of usefulness of the service. Since doctors have the ultimate responsibility for the content of patient notes registered in electronic health record systems, they need to manual fix errors in case they appear there. *"When the transcribers have finished the text, I'll get information: Now the text is available. The rule is that I have to check if it is correct. It's my responsibility." (I2)* This means that quality is important for companies competing for contracts with medical service providers, because the number of billable doctor hours is impacted by the quality of transcripts. Thus, transcribers have a direct impact on competitiveness of their employers. *"We are aiming at perfectionism, so that there is no mistakes. But of course we are human beings […] if you don't hear or understand what the doctor is saying on the audio, you just leave three dots, which indicates that the doctor has to fill it by himself." (I4)* Quality, however, is not the only measure of performance that is of relevance in this case.

Efficiency of transcription, which is the time required to convert audio recorded by a doctor to written patient note, is of critical importance too. In some cases, timely delivery of transcripts can significantly impact patient's long-term health or even make a difference between life and death. *"You have to keep on track [..] and understand really what the doctor is saying. If you type something that doesn't make any sense… it might endanger the patient's treatment." (I4)* Even in less critical situations written patient records are legally required in the Inscripta's country within 5 days from patient's appointment. Thus, speed and efficiency of transcription emerged as one of the central topics discussed by our informants during the interviews.

To track the performance of transcribers Inscripta uses a TTX (Transcription Turnaround Multiple) metric. TTX represents how many minutes it takes for a transcriber to complete, on average, transcription of 1-minute-long audio recording. Thus, reducing TTX indicates task performance improvement in terms of efficiency. Initially, the performance of transcribers at Inscripta has been comparable to that of manual transcribers. *"[…] before the ASR, it [Inscripta's TTX] was somewhere between 4 and 6. And in the public sector, it is between roughly 5 to 8." (I4)* After the introduction of ASR the performance started changing. At first, for some audio recordings the automatically generated transcription was full of errors and transcribers needed to manually type the full text. *"At that time I had to fix the text a lot – there were these really strange terms and sentences" (I6).* However, in some cases the ASR produced good enough results, so that there was a significant reduction in typing, thus leading to overall improvement in TTX. *"The benefit for those where the automatic transcription was going well was already so much that we saw an immediate improvement from 4.5, TTX to 3.5. It went very fast. And then once we updated the model, in no time you're on TTX of two – meaning that we have doubled the efficiency". (I10)* The improvement in TTX continued as transcribers needed to type less and less of the commonly used words and phrases. Eventually, when working with audio recordings from a familiar context, ASR started generating output with a very small fraction of errors. *"We are at the point where with the best dictators we don't have to touch anything anymore. It helps us tremendously." (I9)* Comparing with the initial TTX level the system has improved the performance approximately three times (Figure 2). The subsequent fluctuations resulted from changes in task context, which we cover next.

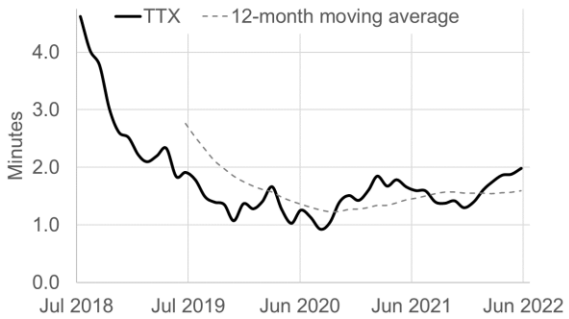

Figure 2. Average time required to transcribe one minute of audio (TTX) at Inscripta.

4.2. Types of context change

Another theme that vividly stood out in our analysis relates to transcription task context and its changes over time. *"I mean, it's an artificial intelligence in the sense that it learns the language, but it doesn't have any kind of worldly knowledge. It doesn't have any kind of contextual information. It doesn't understand what we humans are trying to say, what we want, what we mean by something. So basically, you will still need people to look up whether or not this word actually makes sense in that context. Sometimes it gets it right. Sometimes it gets it horribly wrong." (I8)* Thus, on the one hand the technology part of the system is decoupled from the world and the broader context surrounding the transcription task. It is humans who serve as a bridge between the system and the external context. On the other hand, some part of context is captured by the ML model, and it is leveraged when converting audio to text. *"Context is everything. The same sound […] can mean very different things in different contexts […]. To be able to type in the right word, the computer needs to analyze the context." (D2)* Thus, the functioning of ML algorithm bears some similarity to how people figure out the meaning of sounds. *"For many parts I could reason from the context what the word might be. But there were parts where it was just so obscure that I couldn't reason from the context, what he actually wanted to say" (I3)* This tension between ML algorithms not understanding the broader context yet requiring some of it to function properly leads to another challenge faced by human-in-the-loop systems. Namely, the real-world context is rarely static, thus making the performance of these systems vulnerable to context change. Therefore, considering context changes from the perspective of the system is important.

Over the years, Inscripta experienced three types of context change (or lack thereof): stability, drift, and jump. ***Stability*** is less common and typically relates to doctors that have been served by the company for some time and in the context of, for example, minor cold or flu appointments. Indicators of stability are very high overlap between the ML model training data and new audio recordings that are send by the customers, as well as no changes in the terminology used. *"If you have a lot of examples from that doctor you can actually do very well for that doctor. The words they are saying are surprisingly similar. It's a very structured text. They're not suddenly going to go and talk about whether their dogs got their food this morning, because that just doesn't fit in the medical dictation." (I4)*

Drift is a common type of context change discussed in our interviews. It represents a gradual change, which in case of Inscripta relates to primarily new medical terminology, diseases or names of medicines appearing over time. *"The terminology is never ending and, even today, it expands." (I4)* For instance, changes in the medical terminology due to COVID-19 pandemic constituted a drift. *"COVID-19 is a good example. We didn't have that in our vocabulary prior to year 2019. So those kinds of things you need to add to the vocabulary and there's always some more learning to do." (I8)* Drift might also take place when new individual doctors with distinct pronunciation start using the service. *"If there is a new doctor who emphasizes words or pronounces differently, then you may need to start correcting the text again." (I6)*

Finally, context ***jump*** occurs when there is a sudden and significant change in the content of inputs going into the loop. A jump might arise from external circumstances, although the most prevalent examples in our data relate to changes driven by the company itself. As a growing start-up company, Inscripta continued to expand its service to a broader range of customers. This resulted in contracts with medical clinics and laboratories specialized in medical fields that were new to ASR and, sometimes, also the transcribers. *"We didn't have a customer in the [XYZ medical] field before. The first time when we started doing the dictations with ASR… what it produced in the text… it was basically gibberish. You had to erase it all, and at the beginning it was easier to just type it out completely from scratch." (I9)* Another context jump initiated by the company revolved around serving customers outside of medical field. *"Whenever you move to another discipline the ASR doesn't necessarily improve the efficiency of the transcribers, if, every second word needs to be corrected." (I8)* Finally, Inscripta has been expanding its offering to medical clinics where doctors speak other languages. Change in language represents a major context jump for the loop. Even though there are no changes needed in the setup of the loop or technical architecture, new training data needs to be used. *"For a new language we are training a completely new model. You could transplant an old model to a new language, but as long as you have data, it doesn't make sense to transplant the old one." (I10)*

4.3. Practices for aligning the system with the changing context

Improving transcription task performance and, later, maintaining its good level despite context changes has been critical for Inscripta's growth and commercial viability. Such task performance improvement rests on the ability of humans and technology to learn over time. Human transcribers can type faster as they become more familiar with the terminology. *"You get better at context. When you first start, you might have to look up words or do some research into what is being discussed. You just get faster." (I5)* ML algorithm improves the ability to generate correct words from audio recordings when its outputs are reviewed and, if needed, corrected by the human transcribers. *"The idea is that the ASR should learn after being corrected 10 to 20 times." (I4)* However, Inscripta did not simply rely on uncoordinated learning efforts of humans and technology to improve transcription task performance. Rapid changes in the context and commercial pressures called for a more proactive approach. *"We want to improve the ASR faster than it would be happening if we wouldn't help it. We want to teach ASR as fast as possible and make it as accurate as possible. That's why we do it." (I4)* Therefore, to boost the performance improvement the company has developed two sets of practices for aligning its system with the context – extending and refining.

4.3.1. Extending practices. These organizational practices brought new and relevant pieces of contextual knowledge and understanding from outside of the focal loop. They were carried out by transcribers and other members of Inscripta team, such as technical staff, management, or customer relationship representatives. **Acquiring data from the outside of the system** was very important in the early phases of system development. "*We hired medical students who speak [language ABC] to generate some data for us." (I10) "If you want to recognize social security numbers, you write a small Python script that generates 10,000 variations." (I10)* Furthermore, external data acquisition was also an alignment practice after a context jump, such as new customer contract. *"We always ask customers for audio and text data. If they are willing to share… that would help us a lot, because then we have foundation or base for the field." (I4).* The system relied also on some **hard-coded rules** for proper handling of specific commands. *"We have these lists where we provide the ASR with commands such as start a new line, start a new chapter, etc. And we update them on a regular basis." (I9)* Another important form of extending practice involved **doing background research**. *"Whenever we get a new customer, then one of the more experienced transcribers takes on the vocabulary list of that customer. They figure that out by talking to the customer." (I10)* Some part of background research was more specialized and centered on developing understanding of linguistic issues within the technical team. *"We've been working together and finding how abbreviations [in language ABC] work. I'm doing some research on that for [a colleague from the technical team]." (I9)* Significant share of the background research, however, was allocated to transcribers. *"I would read, for example, something related to [medical field XYZ] or how you prepare a sample. It helps a lot with this work if you're a little bit familiar with the actual work. So, when you hear a new term, you have a vague idea of what it could be. Doing additional reading online around that topic always helps." (I9)* To complement background research transcribers often in the middle of working on a customer job **investigated new cases** of specific terms using medical dictionaries or by searching on Google. *"I want our quality to be good. I instruct our transcribers that every time they encounter a term they haven't seen before, or haven't done research on before, they need to Google that. Because we want to teach the speech recognition to produce the words correctly." (I9)* Finally, management played a role in aligning capabilities and know-how of transcribers with the context by **selecting system users**. This happened via hiring suitable candidates and allocating them to tasks. *"If you have experience in at least one of those medical fields, it's easy to come onboard. Instead, if you have only experience from the basic flu and fever cases, then we'd have to teach you everything." (I4) "If your motivation is to be a part of the healthcare, then transcribing two guys that are chatting about the newest trends [in sector XYZ] might not be motivating. Might it be that this is not going well because, you don't know the [XYZ] terminology and you just find it boring like hell? So, then you work on medical customers and the person who was having more fun with [XYZ]... they can work on that." (I10)*

4.3.2. Refining practices. The second set of context alignment practices propagated and clarified the contextual knowledge and understanding already captured within the system. These practices were equally centered around data handling and creation, as well as continuous learning. However, they focused on refining context alignment and, consequently, task performance by extracting more from what has been already learned and captured by parts of the system. For example, Inscripta's technical team **curated data** used in model retraining to make sure that the most relevant training examples are used. *"In practice, we only add our own data and even make it more important than*

data from other sources." (I10) "According to GDPR you cannot store information indefinitely. For example, after five years, our speech recognizer is starting to suffer from this kind of dementia. So, you need to keep up the service running for it to function properly." (I8) These efforts leveraged the data created by transcribers. Hence, to continuously improve and refine this data, Inscripta developed **quality focus** practices. This included explicitly communicating the importance of quality. *"We have stressed to them that we are teaching our speech recognition and the quality needs to be good. Quality goes first and not the speed." (I9)* Transcribers listened in full to each recording and had to check each unknown word they encounter. *"Even though we have the ASR right now, we always have to listen to the dictation." (I4) "If I get a strange word which looks or sounds weird, then I have to check it. Because I can't live in uncertainty." (I6)* To further strengthen the quality focus, transcribers working at Inscripta had a 6-hour-long workday, while enjoying the same salary levels as full-time transcribers working for competition and had volume targets that were not overly straining them. *"Implementing a 6-hour-workday model that would incentivize quality focus and also give the transcribers their own agency to structure their workday was something that [manager name] was very keen to promote." (D9)* To further capitalize on quality focus, Inscripta continued to **set and enforce consistency standards** that maximize the pace at which ML models can learn. For each new customer they created transcription guidelines and shared them internally. *"There is some manual work involved in making an instruction set for specific clients." (I10)* Also, when transcribers repeatedly encountered different versions of new words, they jointly agreed on spelling conventions for these. *"When Covid came we had to agree about the spelling [of new terms], so that everyone writes the same word the same way. Otherwise, it goes wrong if every transcriber writes it in their own way." (I6)* They were able to promptly detect these new medical terms, because the transcribers communicated with each other regarding unknown words*: "We have this channel in Slack where I guide transcribers on grammar and medical terminology" (I9)* These practices led to both transcribers and technical team **learning from and about the system**. *"When you understand how the ASR is working, it is easier to become even more efficient at using it" (I4)* In some cases this resulted in dealing with problematic inputs by using manual fixes. *"We are realists. Sometimes you have to do the manual fixes. So, we have, a very minimal list of 50 mappings." (I10)* Interestingly, transcribers who work with audio recordings from relatively unfamiliar medical fields learn faster when working along side ASR. *"And then there is the support from ASR. I have noticed that [in XYZ medical field] at the moment it helps a lot. It can predict and know what word should be there. That gives me much more secure feeling." (I6)*

4.4. Synthesis of findings

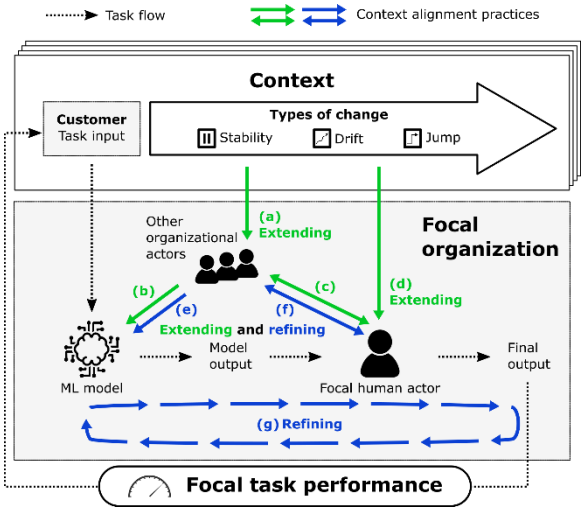

Figure 3. Improving & maintaining performance of a learning human-in-the-loop system via context alignment practices: Extending and refining.

We synthesize the findings into a conceptual model (Figure 3), which combines the insights on how organizational practices related to a learning human-in-the-loop system allow it to improve and maintain performance despite context changes. Inputs for a specific instance of a task originate from a given context. If the context is not static, however, future instances of that task will come from slightly or significantly different context, because of drift or jump, respectively. To align the contextual knowledge and understanding that is captured within the loop with this new and different task context organizational actors engage in extending practices (Table 1A). The pieces of contextual knowledge and understanding, which they bring, are not fully integrated with the loop immediately. It is only the refining practices that propagate and clarify them over time (Table 1B). The combination of these two sets of practices allows the focal organization to improve or maintain high task performance levels regardless of changes.

Table 1. Context alignment practices.

A: Extending	
Definition: Organizational practices related to the focal human-in-the-loop system which bring new and relevant pieces of contextual knowledge and understanding from outside of the loop.	
Individual practices and identified examples:	

• Acquiring data from the outside of the system
  o Getting the training data from third parties
  o Generating new data
  o Augmenting existing data
	a,b

• Doing background research
  o Researching medical procedures
  o Researching linguistic rules
  o Collecting lists of key terminology from customers
	a,c,d

• Investigating new cases
  o Googling
  o Checking medical dictionaries
	d

• Selecting system users
  o Hiring experienced users
  o Matching tasks with interests and/or competences of users
	a,c

• Hard-coding rules o Codifying responses to specific commands	b
B: Refining	
Definition: Organizational practices related to the focal human-in-the-loop system which propagate and clarify contextual knowledge and understanding already captured within the loop.	
Individual practices and identified examples:	

• Curating the training data
  o Prioritizing data from within-the-loop
  o Renewing representative data
	e,g

• Developing quality focus
  o Communicating quality focus
  o Listening through each audio recording in full
  o Requiring validation when an unknown word is encountered
  o Promoting quality focus by giving employees flexibility and achievable targets
	f,g

• Setting and enforcing consistency standards
  o Creating shared instructions for individual clients
  o Agreeing on spelling conventions for new terminology
  o Checking the meaning/spelling of unknown words with colleagues
	f,g

• Learning from and about the system o Getting hints from automatic system outputs	g

5. Discussion and conclusion

Organizations around the world increasingly rely on learning human-in-the-loop systems. Therefore, understanding their impact on task performance is critically important. Scholars have recently started to investigate how such systems are developed (Asatiani et al., 2021; Grønsund & Aanestad, 2020; van den Broek et al., 2020) and what impact they might have on performance and organizational knowledge (Fügener et al., 2021; Ge et al., 2021; Sturm, Gerlach, et al., 2021; Sturm, Koppe, et al., 2021; Sturm & Peters, 2020). The emerging insights, however, do not sufficiently account for context changes that take place in the operating environment of these systems. It is alarming, because both gradual context drift (Kim et al., 2022) and major unexpected jumps that change our world, such as COVID-19, are undeniable elements of the environment in which today's organizations operate. To address this research gap, we have set out to explore how a medical transcription company, which relies on a learning human-in-the-loop system at the core of their operations, improved and maintained performance over a 4-year-long period and despite context changes.

The findings and analysis presented above showed that the performance of medical transcribers working for the case company increased approximately threefold in terms of efficiency once the systems was implemented and developed. The critical insights, however, related to how the organization achieved and maintained this result while expanding their service to cover an increasing range of doctors, medical fields, and clinics, not to mention continuous evolution of medical vocabulary and disruptions caused by COVID-19. To deal with these context drifts and jumps, the company relied on two sets of **context alignment practices**. By employing **extending practices,** the employees brought into the learning loop new and relevant pieces of contextual knowledge and understanding from the outside. Thus, extending practices served as a bridge connecting the loop with the external world. Next, **refining practices** enabled and facilitated the propagation and clarification of the contextual knowledge and understanding that was already captured within the loop. These refining practices ensured that ML algorithm continued to learn quickly and achieved high level of accuracy. Simultaneously, these practices resulted in human agent learning.

Insights from our work advance the emerging IS and organizational literature concerned with learning human-in-the-loop systems (Baird & Maruping, 2021; Fügener et al., 2021; Ge et al., 2021; Grønsund & Aanestad, 2020; Lyytinen et al., 2020). By focusing on task-level changes, we respond to Raisch and Krakowski's (2021) call for conducting micro-level research in the area. We also follow a recent plea to develop deeper insight into the critical organizational functions underlying the organizational design and integration of these systems while recognizing the sociotechnical nature of the underlying phenomena (Benbya et al., 2021; Lyytinen et al., 2020).

Our first contribution is showing the nuance of how context changes influence learning human-in-the-loop systems. Performance deterioration due to gradual context drift is a well-recognized challenge in machine learning applications (Kim et al., 2022). Based on our findings, it is distinctly necessary to recognize context jump. When a jump occurs the focal task and the type of inputs it receives remain fixed, but the content of these

inputs dramatically changes over a short period of time and, thus, greatly undermines ML performance. Our work in this respect complements and goes beyond the recent research in three notable ways. First, we vividly show the importance of understanding the sociotechnical aspects of organizational approaches to dealing with context changes. Thus far the emphasis has been on technology-focused methods (Kim et al., 2022), which we view as only one part of a greater solution. Second, it is necessary to study the performance of human-in-the-loop systems across the full range of changes and not just stability (Fügener et al., 2021), drift (Sturm, Koppe, et al., 2021), or jump (Grønsund & Aanestad, 2020) in isolation. Only by understanding whether and how organizations manage human-in-the-loop systems across these three can we draw more generalizable performance conclusions. This is especially important because context is rarely stable or drifting slowly forever. Third, our results help with interpretation of recent conceptual work recognizing the importance of considering context in humans-ML collaboration (Baird & Maruping, 2021). Contrary to Balasubramanian and colleagues (2020, see proposition 3a), we find that both in moderately (drift) and highly (jump) changing context human and machine collaboration delivers superior performance and learning, when compared to only human or only machine setup. Also, counter to their explanation, our results indicate that high rates of change (jump) require substantial human involvement to maintain or improve system performance and learning. Thus, our findings concur with the view that in turbulent environments humans need to actively engage in exploration to seek out new knowledge and collaborate with each other to efficiently and effectively train the ML algorithm (Sturm, Gerlach, et al., 2021, see proposition 3).

The second set of theoretical implications of our study revolves around the fine-grained understanding of organizational practices employed in improving and maintaining performance of a human-in-the-loop system. Notably, we found that the context alignment practices of extending and refining are not simply technical procedures, hyperparameter settings, or algorithms. They do not take place in organizational vacuum. Instead, they involve more complex interactions happening at the sociotechnical system level (Mucha et al., 2022; Winter et al., 2014). This finding aligns with those of Asatiani and colleagues (2021). They conclude that the technical elements of envelopment enabling safe and socially responsible usage of an ML-based application intersect with social factors, thus making human agents an integral part of the practices. In our case, human agents actively communicate with each other when enacting the context alignment practices. This means that we question whether studies investigating the performance of human-in-the-loop systems, where human actors do not coordinate or communicate with each other (Fügener et al., 2021), can be generalized to organizational setting. Furthermore, our findings show a comprehensive set of practices needed to improve and maintain the performance of learning human-in-the-loop systems. Thus, we go beyond the insights from the case study carried out by Grønsund and Aanestad (2020). They describe how different human roles emerge and develop, and recognize two high-level types of algorithm augmentation work: auditing ("i.e., the generation of a ground truth and assessment of the algorithmic output against this") and altering (i.e., "the work of altering the algorithm and the data acquisition architecture"). Based on these and our findings we conclude that system learning capacity and human agents' exploration work are deeply interconnected. Hence, for human-in-the-loop systems we refute the notion that ML algorithms reduce the need for human exploration (Sturm, Gerlach, et al., 2021, see proposition 1).

The insights presented here are equally relevant for practitioners. Managers can use the practices we identify to inform the design and development of learning human-in-the-loop systems. We demonstrate that such systems need a set of integrated human roles and practices that enable and facilitate performance improvement and retention in changing context. Whether the change is triggered by incremental drift or major jumps the organizations need the capability to align the loop with context by jointly relying on extending and refining practices.

To conclude, we see the present case study as a spark that might ignite further research efforts to better understand the role and impact of context changes on the human-in-the-loop systems and their constituent sociotechnical systems. This study is subject to several limitations that invite further exploration. The case company was a start-up organization and the focal task, the generation of medical transcriptions, had an ideal set-up for learning human-in-the-loop system, because the ground truth has been relatively easy to identify and there were no delays between task-related actions and outcomes. The low ambiguity, relative ease of generating labelled training dataset, and economics that support full supervision of the system by transcribers make this case a suitable reference point to compare future case studies dealing with more complex settings.

6. References

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2021). Sociotechnical Envelopment of Artificial Intelligence: An Approach to

Organizational Deployment of Inscrutable Artificial Intelligence Systems. *Journal of the Association for Information Systems*, *22*(2), 28.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on Is Use: A Theoretical Framework of Delegation to and from Agentic Is Artifacts. *MIS Quarterly*, *45*(1), 315–341.

Balasubramanian, N., Ye, Y., & Xu, M. (2020). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. *Academy of Management Review*.

Benbya, H., Pachidi, S., & Jarvenpaa, S. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, *22*(2), 10.

Berente, N., Bin Gu, Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3).

Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, *71*.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4).

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with Ai. *MIS Quarterly*, *45*(3), 1527–1556.

Ge, R., Zheng, Z. (Eric), Tian, X., & Liao, L. (2021). Human–Robot Interaction: When Investors Adjust the Usage of Robo-Advisors in Peer-to-Peer Lending. *Information Systems Research*, *32*(3), 774–785.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*.

Graebner, M. E., Martin, J. A., & Roundy, P. T. (2012). Qualitative data: Cooking without a recipe. *Strategic Organization*, *10*(3), 276–284.

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, *29*(2), 101614.

Hu, W.-L., Rivetta, C., MacDonald, E., & Chassin, D. P. (2019). Optimal operator training reference models for human-in-the-loop systems. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology*, *56*(1), 517–543.

Kim, S., Mai, T.-D., Han, S., Park, S., Nguyen, T., So, J., Singh, K., & Cha, M. (2022). Active Learning for Human-in-the-Loop Customs Inspection. *IEEE Transactions on Knowledge and Data Engineering*, 1–1.

Koppe, T., & Schatz, J. (2021). Cloud-based ML Technologies for Visual Inspection: A Case Study in Manufacturing. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 1020.

Lyytinen, K., Nickerson, J. V., & King, J. L. (2020). Metahuman systems = humans + machines that learn. *Journal of Information Technology*.

Mucha, T., Ma, S., & Abhari, K. (2022). Beyond MLOps: The Lifecycle of Machine Learning-based Solutions. *AMCIS 2022 Proceedings*, 11.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3).

Paterson, C., Calinescu, R., & Ashmore, R. (2021). Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Computing Surveys*.

Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review*, *46*(1).

Ransbotham, S., Khodabandeh, S., Kiron, D., Candelon, F., Chui, M., & LaFountain, B. (2020). Expanding AI's Impact With Organizational Learning. *MIT Sloan Management Review and Boston Consulting Group*.

Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity. *Journal of the Association for Information Systems*, *22*(2), 9.

Sturm, T., Gerlach, J. P., Pumplun, L., Mesbah, N., Peters, F., Tauchert, C., Nan, N., & Buxmann, P. (2021). Coordinating Human and Machine Learning for Effective Organizational Learning. *MIS Quarterly*, *45*(3), 1581–1602.

Sturm, T., Koppe, T., Scholz, Y., & Buxmann, P. (2021). *The Case of Human-Machine Trading as Bilateral Organizational Learning*. 18.

Sturm, T., & Peters, F. (2020). The Impact of Artificial Intelligence on Individual Performance: Exploring the Fit between Task, Data, and Technology. *ECIS 2020 Research Papers*, 17.

Sun, H., Wright, R. T., & Thatcher, J. (2019). Revisiting the impact of system use on task performance: An exploitative-explorative system use framework. *Journal of the Association for Information Systems*, *20*(4), 3.

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-Ml Augmentation. *MIS Quarterly*, *45*(3), 1483–1499.

van den Broek, E., Sergeeva, A., & Huysman, M. (2020). Hiring algorithms: An ethnography of fairness in practice. *ICIS 2019*, 1–9.

van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the Machine Meets the Expert: An Ethnography of Developing Ai for Hiring. *MIS Quarterly*, *45*(3).

Winter, S., Berente, N., Howison, J., & Butler, B. (2014). Beyond the organizational 'container': Conceptualizing 21st century sociotechnical work. *Information and Organization*, *24*(4), 250–269.

Yin, R. K. (2009). *Case study research: Design and methods* (Vol. 5). sage.