Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants

CARRIE J. CAI, Google Research, USA
SAMANTHA WINTER, Google Health, USA
DAVID STEINER, Google Health, USA
LAUREN WILCOX, Google, USA
MICHAEL TERRY, Google Research, USA

Deep neural networks (DNNs) routinely achieve state-of-the-art performance in a wide range of tasks, but it can often be challenging for them to meet end-user needs in practice. This case study reports on the development of **human-AI onboarding materials** (i.e., training materials for users prior to using an AI) for a DNN-based medical AI assistant to aid in the grading of prostate cancer. Specifically, we describe how the process of developing these materials changed the team's understanding of end-user requirements, leading to fundamental modifications in the development and assessment of the underlying machine learning model. Importantly, we discovered that onboarding materials served as a useful **boundary object** for cross-functional teams, uncovering a new way to assess the ML model and specify its end-user requirements. We also present evidence of the utility of the onboarding materials by describing how it affected user strategies in a study deployment to pathologists.

Additional Key Words and Phrases: datasets, neural networks

ACM Reference Format:
Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *CHI 2021.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Although deep neural networks (DNNs) routinely achieve state-of-the-art performance in a wide range of tasks, it can often be challenging for these AI-powered systems to meet end-user needs in practice [3, 14]. This case study describes how the process of developing **human–AI onboarding materials** (i.e., training materials for users prior to using an AI) for a medical AI assistant altered the development team's understanding of the AI's end-user requirements, leading to changes in the development and assessment of the underlying machine learning (ML) model. As such, the onboarding materials served as an important facilitator and tangible glue between user needs and ML model development.

Approximately 12% of men will be affected by prostate cancer in their lifetime [1]. Prostate cancer grading involves a combination of detecting, classifying, and quantifying the amount of different cancer severity grades, making it a highly challenging task that can often lead to inter-rater reliability. Recently, it has received significant attention in ML-powered cancer diagnosis [5, 9]. To aid the accurate detection and grading of prostate cancer, we developed an ML model to identify and grade the cancer in prostate tissue biopsies [9]. The intent of the model is to increase the

Manuscript submitted to ACM

consistency and accuracy of this task. To assist pathologists, the model provides the core functionality for an AI assistant that is accessible within a digital slide viewer. The slide viewer displays tissue biopsies, with the AI assistant overlaying its localized predictions on the tissue (users can hide predictions, and can also perform operations such as pan and zoom). The intent of the AI assistant is to aid the pathologist in their cancer grading task, rather than replace them, and was developed through an iterative design process informed by many rounds of formative studies with pathologists.

In our early work [4], we conducted a series of interviews with pathologists to determine what information they thought should be included in training materials to effectively use the AI assistant during the diagnostic process. The interview results not only informed the design of onboarding materials for the prostate cancer model, but also established **onboarding guidelines** for introducing end-users to medical AI assistants in general [4]. According to these guidelines, critical information about an algorithm should be presented to human experts upfront, prior to use, for the purpose of developing appropriate mental models and strategies of use. For example, pathologists in our study needed to know about the model's strengths and limitations, the model's source of ground truth (e.g., which physicians provided ground-truth labels for the training data), and the types of idiosyncratic, AI-specific errors the system could make that would be unexpected to human experts [4].

While that prior work established a theoretical framework for onboarding, the current case study uncovers the practical realities of implementing the onboarding guidance. Through our process of developing onboarding materials for the prostate cancer grading, we found that many types of information required for human–AI onboarding were challenging to immediately fulfill in practice, due to the way the AI model had been designed, and the fundamental assumptions shaping the model development and data labeling process. This mismatch led to discussions and rounds of iteration between cross-functional teams, ultimately resulting in changes to the model's requirements and development process that we outline below.

Importantly, through this experience of discussions and iteration, we discovered that the creation of onboarding materials both changed the team's understanding of end-user requirements for the AI, and uncovered new ways to assess the underlying ML model and specify its end-user requirements. Specifically, the onboarding information requested by pathologists initiated new efforts to:

- Broaden model evaluation to include a systematic analysis of *pathologists'* strengths and weaknesses vis-a-vis the model, to determine in which sub-scenarios users could most benefit from AI assistance.
- Deepen model evaluation beyond aggregate performance metrics, to include a stratified assessment of the model along *domain-specific cases* (e.g., "How well does the model handle tissue that is crushed?"), to improve end-user understanding of categorical model strengths and limitations.
- Modify data labeling pipelines to incorporate labels of domain-specific cases (e.g. whether the tissue is crushed), beyond acquiring labels optimized for model performance (e.g., cancer grade classification).
- Collect and centralize distributed engineering knowledge about model idiosyncrasies that may diverge from human idiosyncrasies.
- Translate model metrics into human-digestible takeaways.
- Collectively define the model's objective function within the broader diagnostic workflow.

These results suggest that the onboarding materials served as a useful **boundary object** between cross-functional team members. Boundary objects [6, 7, 10] are defined as common frames of reference that enable interdisciplinary teams to cooperate, despite having different methods, data abstractions, time horizons, and audiences to satisfy. In our context, onboarding materials further acted as critical *boundary negotiating artifacts* [7], providing a concrete means for

HCI researchers, machine learning developers, product managers, designers, and end-users to discuss and converge on a final AI system design enabling effective human–AI collaborations. These findings echo experiences of others who have found that writing user manuals can help establish a "common ground" between stakeholders [12]. These onboarding materials were ultimately deployed as part of a larger machine learning study [11] intended to test the AI model's efficacy in assisting pathologists. Through feedback on the deployment, we found that onboarding played a critical role in defining users' decision-making strategies and approaches when partnering with the AI assistant.

In the rest of this case study, we first provide a background summary of human–AI onboarding guidelines, then reflect on how the practical implementation of onboarding changed the course of model development, acting as a boundary object (and boundary negotiating artifact) between cross-functional stakeholders. Finally, we describe how the onboarding materials were ultimately used and perceived in a study deployment to pathologists.

2 BACKGROUND

As mentioned in the Introduction, our prior work identified a set of information pathologists considered useful for learning how to effectively use a medical AI assistant [4]. This information includes:

- **Capabilities and limitations**: How well the Assistant performs overall, as well as how it performs in domain-specific contexts that humans find difficult.
- **Functionality**: The information the Assistant uses in making its recommendations (e.g., does it have access to the patient's history).
- **Medical point-of-view**: The AI assistant's source of ground truth (who provided the ground-truth labels for the data set), and how the Assistant is biased (e.g., is it more conservative or liberal in grading cancer severity).
- **Design objective**: What objectives the Assistant has been optimized for (e.g., standalone vs human-collaborative use, sensitivity vs. specificity, or whether it was intentionally designed to compensate for human errors).

We applied these guidelines to the design of the onboarding materials, used as part of a larger model deployment study examining the effectiveness of the AI assistant [11] for prostate cancer grading. In the study deployment, 20 pathologists reviewed prostate needle biopsies with and without AI assistance, quantifying and grading the severity of cancer (e.g. Gleason grade group 1, grade group 2, grade group 3, etc.). The onboarding materials we developed for the study included instructions on how to use the interface (e.g., available features, keyboard shortcuts) as well as information describing:

- The amount and type of training data used to train the model, and the inputs to the model (what information it does and does not take as input).
- The qualifications of the pathologists who annotated the training data (source of ground truth).
- The strengths and weaknesses of the Assistant (e.g., in which cases it has high concordance with experts, and the types of cases that can induce errors), vis-a-vis the strengths and weaknesses of general pathologists.
- Strategies for using the Assistant in light of its known strengths and weaknesses.

We provided the onboarding materials in a digital slide deck, which participants reviewed prior to starting the experiment. Participants were also provided a short summary of the onboarding material each week of the study to help refresh their memory (there were a total of eight weeks in which participants reviewed slides). At the end of the study, participants responded to a post-study survey that included questions about the onboarding materials.

The model deployment study found the AI-based assistance to significantly increase participants' diagnostic performance (agreement with subspecialists' grading), to reduce review time, and to improve pathologists' self-reported

confidence [11]. In the current case study, we specifically focus on the practical realities of creating and deploying onboarding in the broader cross-functional process, and its impact on end-user strategies when collaborating with AI.

3 ONBOARDING AS BOUNDARY OBJECT BETWEEN HCI AND MODEL DEVELOPMENT

In developing the onboarding materials, we found that the information requested by pathologists for the onboarding materials broadened the team's understanding of how the pathologists assessed the model and what features they considered important to be able to effectively use it in practice. Collecting this information also influenced the research process itself. The subsections below (summarized in Table 1) describe the impact specific information needs had on the research, the team, and its practices.

3.1 Systematic Identification of User Strengths and Weaknesses (vis-a-vis the AI)

One common request of pathologists was to understand the model's strengths and weaknesses. However, we found that simply providing common performance metrics (such as precision and recall) was inadequate: users did not know how to transform these metrics into actionable strategies for partnering with the AI. One reason for this difficulty is that pathologists did not know how the model's performance compared to their own strengths and weaknesses as pathologists. Without an awareness of their own relative performance, broken by down specific types of cases, they were unable to strategize in which scenarios they should rely on the model, and in which scenarios they should rely on their own judgment.

Meanwhile, the engineering team had started benchmarking model performance against the diagnostic performance of general pathologists (unassisted), to measure model efficacy. This analysis revealed that, across the four grade groups, one area of substantial benefit for the model was in accurately distinguishing between grade groups 1 and 2. Building on this finding, we identified that the analysis being conducted for model evaluation could be expanded to develop and communicate a decision strategy to end-users during onboarding.

While the benchmarking analysis produced data comparing pathologist and model performance (stratified by cancer grade), we found that simply showing these metrics to users was insufficient for helping them create a mental strategy for working with the AI. For example, beyond accuracy numbers, pathologists were often concerned with the possibility of over-grading (grading the severity of cancer too high) or under-grading (grading the cancer severity too low). Moreover, whereas accuracy metrics typically capture performance relative to ground truth, pathologists found it more actionable if metrics were framed relative to *their own diagnostic instinct* (e.g. "*If I think this case is grade group 2*, how likely is it that I would be over-grading? How likely is it that the model would be over-grading?"). Taken together, these analyses helped the team establish one important advantage of the model to be in helping pathologists not over-grade grade group 1 cancer as grade group 2.

In onboarding materials, we included an explicit user strategy describing potential cases in which to carefully consider the model's prediction. Notably, given the user feedback above, we framed the strategy with respect to the user's own initial diagnostic instinct: *"If you are unsure between grade-group 1 and grade-group 2, AND you see that the AI assistant is calling grade-group 1, this might be a scenario to carefully consider the AI's suggestion in order to avoid overcalling grade-group 1 as grade-group 2."* This strategy was noted in post-study feedback as being the most useful piece of content from onboarding. This information may also help engender appropriate levels of trust in the AI [13, 15] by illuminating in which situations they should pay special attention to the AI's predictions.

Table 1. Onboarding Information Requests and Impact

Onboarding: User information requested	**Impact of collecting desired information**
Actionable strategy for when to rely on AI vs. when to rely on own judgment	**Discovery**: User (i.e., pathologists') weaknesses were unknown (i.e., types of cases they are less likely to accurately grade). **Result**: Expanded comparative analysis of pathologist vs. ML performance. Moved beyond typical performance metrics towards data insights that are actionable to end-user's frame of reference (e.g., *"If I think a case is cancer grade 2, how likely am I over-grading vs. the model is over-grading?"*).
Human-meaningful model strengths and weaknesses (e.g. performance on domain-specific subtypes / edge-cases)	**Discovery**: Because upstream data labeling had been optimized for model performance (with coarse-grained classification labels), collecting more detailed sub-type labels would be costly. Discovered that sub-type labeling could serve a dual purpose—not only in setting user expectations of model limitations during onboarding, but also in continuing to improve the model itself. **Result**: Data relabeled with sub-types (e.g., sub-patterns of cancer grade 5; cases with tissue artifacts) to improve both ML and onboarding, in parallel.
"Non-human" model idiosyncracies and unexpected behavior	**Discovery**: Knowledge about model idiosyncrasies were distributed across engineering team members, making it difficult to convey pattern-matching idiosyncrasies of modern machine learning. **Result**: Prompted the collation of distributed knowledge among engineering team members about model idiosyncrasies. **Result**: Stimulated the development of introductory materials for AI so pathologists can form a more accurate mental model of how it works (e.g., visual pattern-matching, lack of explicit training on biological concepts, the fact that its predictions will look "blocky" due to its implementation).
Translating model information into human-digestible takeaways	**Discovery**: Model metrics, graphs, and onboarding checklists were insufficient for translating model information into human-digestible takeaways. A "translation" step was necessary. **Result**: Prompted the engineering team to reframe model properties and metrics with respect to the end-user's theory of mind. Necessitated co-creation of onboarding between engineers, project managers, UX designers, and doctors (end-users).
Model's design objective	**Discovery**: Mismatches between user needs and model objectives (e.g., trade-offs between sensitivity vs. specificity, differential cost of error types, ability to compensate for common human errors, etc.) **Result**: Prompted a shift in model objectives: engineering team augmented data annotation to re-train the model on a subset of cases deemed particularly tricky to pathologists (e.g., missed tumor, artifacts, pre-cancerous entities).

In sum, pathologists' requests for specific strategies in using the AI assistant not only resulted in enhanced onboarding materials, it also granted the team a deeper understanding of the specific and unique strengths of the tool they were building.

3.2 Revamping Data Labeling Pipeline to Include Human-meaningful Subtypes

Beyond the high-level guidance of how to use the Assistant, pathologists also wanted to know how the model performed in specific contexts. For example, in cancer diagnosis, there are specific subtypes and edge cases that doctors tend to find particularly difficult to diagnose (e.g. cribriform pattern, biopsy artifact). For these specific subtypes, the pathologists wanted to know the model's behavior so they could further develop their strategy for when to rely on the model versus their own judgment.

While pathologists wanted to know the model's behavior on these subtypes, we initially lacked these data—subtype labels were not being collected as part of the data labeling process (the model's primary task is to predict the Gleason grade, so data labeling was optimized for collecting Gleason grade labels). Without the desired subtype labels, we were unable to immediately compute model performance metrics on subtypes. Labeling medical data is also resource intensive, as it takes significant time and effort on the part of experts. As such, we did not immediately dedicate resources to collecting subtype labels to address the information request for the onboarding materials.

During model development, improvements to model performance started to plateau. At this point, the team hypothesized that identifying and targeting the model's "error modes" could potentially lead to improvements in the overall model performance. Accordingly, labeling efforts were expanded to include labeling specific subtypes. While these additional labels were primarily collected to improve model performance, they also enabled us to compute the data necessary for communicating model strengths and weaknesses during onboarding.

A number of observations can be made from this experience of identifying and labeling subtypes known to be problematic for experts. First, identifying the specific contexts in which people already have difficulty is a useful step, as it suggests specific areas of focus in the design and evaluation of the tool. Second, these specific contexts suggest areas in which the AI should be evaluated and critically examined. While aggregate performance metrics provide an overview of the model's capabilities, it may be the case that the model is performing well in situations where the typical user will also perform well, but performing poorly in cases where the typical user also performs poorly. Identifying these challenging contexts helps the team assess the model in user-critical situations, and identify circumstances where model errors may be tolerated by end-users. Third, communicating these error modes to end-users helps set expectations about model performance, potentially avoiding frustration and erosion of trust should the errors be encountered during tool use. Finally, knowing what these challenging contexts are can inform the data labeling process by indicating what types of examples should be included in the dataset, to ensure the model can learn to act appropriately in those circumstances. Interestingly, we found that the labeling of human-meaningful subtypes served a dual purpose: not only did it help set user expectations about model strengths and weaknesses during onboarding, but it also helped improve the model. In future efforts of this kind, keeping this duality of purpose in mind may help HCI and ML teams in aligning their efforts.

3.3 Assembling Distributed Knowledge about Model Idiosyncracies

Roughly speaking, the image model developed for this work operates by learning associations between visual patterns and human-provided labels. As such, it has no other explicit information to utilize besides visual patterns (e.g., it was not explicitly trained on biological processes such as cell growth or proliferation). Pathologists, on the other hand, have the additional context of the underlying biological and clinical processes. Devoid of this additional context, a model

may occasionally make predictions that seem bewildering to a pathologist. For example, the model may arbitrarily divide a prediction across a biological structure, such as a single gland, and indicate that one half is cancerous while the other half is not. A pathologist would never make these types of segmentation mistakes, leading a pathologist to question how much they can trust the model, even if the model can be relied on in most other cases. For this reason, it can be useful to explicitly call attention to these "non-human" AI behaviors during onboarding to properly calibrate user expectations and improve model trust.

In cataloguing these model idiosyncrasies, we found that knowledge of these types of model mistakes was distributed among many different team members: Since different people were developing different parts of the model, there wasn't a single place where model idiosyncrasies were being globally tracked. As a result, the creation of onboarding involved collating and synthesizing anecdotal knowledge across multiple team members.

As with the other types of information gathered for onboarding, understanding the idiosyncracies of the model behavior is not only beneficial for inclusion in training materials, it also helps to raise awareness of the end-users, their needs, and their mental models. Accordingly, it may be a worthwhile practice to regularly note these types of idiosyncratic behaviors throughout the development process, not only so they can be more easily incorporated in onboarding materials later, but also to remain sensitive to the end-user's perspective.

3.4 Translating Model Information into Human-Digestible Takeaways

To help other teams produce onboarding materials, we created a list of questions for ML engineers to complete about the models they are building. In testing these questions with another team, we found that engineers were able to answer most questions, but the answers did not always feel actionable to end-users in terms of building a strategy for partnering with the model.

As an example, in response to a question asking for a description of model inputs, an engineer responded, "images and metadata," but wasn't otherwise sure how much detail to go into with regard to the metadata. Accordingly, we reworded the question to be relative to an end-user's theory of mind [2]: "What inputs does it *not* take in that a doctor would normally consider? What inputs does it take in that aren't shown to the user?" This rephrasing led to much more specific and informative content. For example, the engineer subsequently described information that the model was not trained on (e.g., disease progression over time), and also framed model inputs with respect to whether all inputs could be seen in the user interface (e.g., "the doctor and the model should see the exact same thing [on the page]"). We integrated these re-phrased versions into the onboarding materials.

We similarly observed that it can be difficult for team members to describe AI idiosyncrasies in a way that would make sense to end-users. In part, this difficulty is due to the challenges in explaining the pattern-matching aspects of modern ML, and how this pattern-matching might produce biologically implausible outcomes. Ultimately, we found it most effective to directly explain types of ML behaviors and idiosyncrasies to users: *"Rather than being explicitly trained on textbook biology concepts, the model learns to associate visual patterns with diseases, which could lead to strange behaviors when seen from a clinical or biological perspective. If you see obvious errors, these errors may make more sense if you remember that it is trained on visual patterns."* Additionally, we found it was important to cement this high-level idiosyncrasy with one or two concrete examples of bewildering errors the model might make due to pattern-matching.

3.5 Jointly Defining Model Objectives

In some cases, the onboarding creation process also enabled teams to discover mismatches between user needs and model objectives. For example, pathologists reacted far more strongly to model false negatives than to false positives:

Whereas false positives could presumably be caught by the pathologist and reviewed, false negatives (e.g., missed tumor) were deemed more harmful because they would not otherwise be brought to the pathologist's attention. The model was subsequently tuned to place a greater emphasis on not missing patches of tumor.

Users also noted that certain cancer grade errors were more costly than others, due to some grades having a more pivotal effect on patient prognosis and treatment. For example, some felt it particularly alarming if the highest prostate cancer severity (Gleason pattern 5) was over-called. In addition, some users also asked if the model could focus its efforts on particular subsets of patterns that they themselves found particularly difficult to categorize (e.g., tricky histologies such as artifacts and pre-cancerous entities). The team subsequently enriched data collection for those types of cases to optimize the model.

4 ONBOARDING USAGE AND IMPACT ON HUMAN–AI DECISION-MAKING

As we noted above, producing the onboarding materials had the useful side effect of deepening the team's understanding of users, their requirements, and the overall strengths of the tool being built. The model deployment study in which onboarding was used provided additional data about how end-users apply the onboarding materials in practice.

At the conclusion of the study deployment, participants were asked whether and how they used information from onboarding when reviewing cases assisted by AI [11]. In addition, they were asked to describe which aspects of the onboarding material were most memorable when using the AI, and to reflect on how their decision-making could have been different had they not seen the onboarding material. 17 participants completed this post-study feedback (3 were excluded from this analysis due to non-completion).

Many participants indicated that the most valuable takeaway from onboarding was the provision of an actionable strategy (shown in Section 3.1 paragraph 4) for partnering with the AI when their opinions conflict. Specifically, they appreciated knowing on which types of cases the AI is likely to be more accurate compared to a pathologist: *"I found this statement the most useful and kept it in mind throughout the study."* Several participants described how they actively applied this strategy when making difficult decisions in partnership with the model: *"Due to onboarding, I also reconsidered my diagnosis...if the Assistant predicted 3+3"*; *"In cases where I questioned whether the tumor is 3+3 vs 3+4, I went back to this background information and my decision on final grading was influenced by knowing this data."* (Note that "3+3" refers to prostate cancer grade group 1, and "3+4" refers to prostate cancer grade group 2.)

Some participants also indicated that onboarding helped them become more aware of their own limitations, such as human tendencies to over-grade in certain situations: *"The...thing that stood out was that I should try not to overdiagnose 3+4."* As such, onboarding helped hone their attention towards situations where they should slow down and more carefully consider the AI's prediction: *"I...remembered to really pay attention when the Assistant called 3+3 and I was contemplating 3+4."*

Aside from actionable strategies, many also expressed the importance of knowing that the model had been tested against a groundtruth of labels provided by urologic subspecialist pathologists, rather than general pathologists: *"I may have been more apt to ignore the model for my own initial impression if not given more background on how well the model agreed with sub-specialists versus general pathologists."* As such, knowing that the source of ground truth was based on the judgment of highly qualified experts appeared paramount to initial impression formation: *"The assistant's diagnostic capabilities compared to that of GU experts gave me more confidence in the case findings."* A few participants felt that the onboarding explanations about model limitations and idiosyncrasies helped calibrate their expectations, making it less startling when they encountered these algorithmic error modes during diagnosis: *"Was helpful to...know that artifacts would sometimes be highlighted erroneously."*

Finally, some users felt that, without onboarding, they could have made inaccurate assumptions about the Assistant: *"I think I would have partially misunderstood [the] Assistant and its abilities."* These assumptions could in turn lower trust in the model: *"My level of trust in the Assistant grades would have been lower without the background information presented in the onboarding material."* Without the strategies presented in onboarding, participants felt it would be more difficult to decide what to do in cases where their opinion conflicted with that of the model: *"It would have been more difficult and worrisome when my diagnosis conflicted with [the Assistant]."* Others also felt that it would have *"taken more time to figure out the strengths and limitations of the Assistant,"* and that *"it would have taken some time to adjust to the new tech. Onboarding made it seamless."*

We also saw some evidence that it may be difficult for users to remember and retain key information shown in onboarding: *"It was hard to absorb all of the onboarding material without seeing how the assistant tools are actually used."* Even though participants were provided a weekly onboarding refresher, in post-study interviews, nine out of 17 participants brought up the actionable strategy, and eight did not. It is possible that some participants did not review this refresher on a weekly basis: *"I probably should have reread the onboarding material after using the tools for a couple of weeks."* To this point, some felt it helpful to bookmark the onboarding summary in their browser, so that they could revisit it whenever they needed a refresher. These findings suggest that, beyond onboarding users prior to the use of AI, it may be equally critical to embed onboarding refreshers directly into end-users' long-term workflows, so that they regularly encounter the information as it is needed.

The actionable strategy in onboarding (regarding the specific scenario in which to more closely consider the AI) appeared to be pivotal to decision-making approaches. However, only half of the participants explicitly brought up this strategy when asked to recall valuable takeaways from onboarding. Given this, we compared the diagnostic performance in the study of those who had recalled the onboarding summary and those who did not. We found that, in the AI-assisted condition, those who recalled the strategy had significantly higher accuracy when grading the cases compared to those who did not ($\mu = 78\%$ (recalled strategy), $\mu = 72\%$ (did not recall), $p = 0.003$, $t = 3.7$). No difference in performance was found in the non AI-assisted condition ($\mu = 71\%$ (recalled strategy), $\mu = 69\%$ (did not recall), $p = 0.48$, $t = 0.73$). Furthermore, those who recalled the strategy appeared to have benefited significantly from being AI-assisted compared to no-assistance ($\mu = 78\%$ (AI-assisted), $\mu = 71\%$ (no assistance), $p < 0.0001$, $t = 5.5$). While this analysis was conducted post-hoc, and causation cannot be concluded, these findings corroborate the qualitative feedback that onboarding may have played a pivotal role in producing more effective human-AI partnerships.

5 DISCUSSION AND IMPLICATIONS

In this case study, we found that the process of producing onboarding materials served as a means for cross-functional team members to jointly understand how the AI system would be perceived and assessed by end-users in practice, leading to changes in model development and model objectives. While creating onboarding materials, we found that upstream model development and data-labeling decisions had significant downstream consequences, restricting what information could be provided in onboarding. Because these realizations came after the model had already been partially developed, the team incurred some overhead in revamping upstream model development, relabeling training data, conducting additional evaluations to identify human strengths and weaknesses, and redefining model objectives. This brings to light an age-old chicken-or-egg problem in developing ML models intended to assist users in a task: realistic user testing cannot be conducted until a model is built, yet model development depends on end-user requirements.

These observations suggest the potential utility of **developing onboarding materials early in the design and development process**, to align cross-functional stakeholders and raise awareness of end-user needs and requirements

before a model is built. Despite the existence of early testing methods (e.g. Wizard-of-Oz), it is currently still challenging to collect user requirements in abstraction, before a model has been built, in a way that can be made tangible to both HCI and ML teams. Onboarding has the potential to provide a more concrete, tangible means for quickly "testing out" a model's objectives before it is fully built, not only in its presentation to end-users, but also in its role as a joint frame of reference for diverse stakeholders.

This raises a question of which onboarding components can be leveraged early in the model development process to align initial efforts, and which components require knowledge of the actual model to be gathered later in the process. For example, a systematic analysis of user strengths and weaknesses (Section 3.1) could be identified independent of the model, presented to end-users in an "as-if" onboarding to demonstrate model objectives (e.g. sub-areas in which the model could be particularly helpful), and finally used to focus the scope of data collection and model training. Conversely, specific model strengths and weaknesses (Section 3.2) may be more difficult to first identify prior to the model being built. However, there may be a common set of model idiosyncrasies (e.g. ML's visual pattern-matching nature) (Section 3.3) that can be reused from previous models, and used to seed initial onboarding materials for a new end-user audience. Overall, we suggest using onboarding as a formative, cross-functional exercise to identify specific model objectives based on human strengths and weaknesses, then iteratively refining onboarding materials (e.g. to flesh out model idiosyncrasies) as the model develops.

While our onboarding development process was organic and opportunistic, our experience also suggests new avenues for **better codifying this onboarding creation process**, so that it can be reused across teams and encouraged earlier in the development process. For example, just as model specifications, business justifications, and design documents are often produced early in a product life cycle, onboarding materials (with its accompanied user feedback) could be made a tangible requirement for early check-off. Our ongoing efforts suggest that an early, time-bounded session during which cross-functional stakeholders observe an end-user's reactions to onboarding could go a long way towards jointly defining and redefining model objectives.

To our surprise, in-depth descriptions and graphs of model performance metrics were largely insufficient for providing users an *actionable strategy* for partnering with the model: a translation step was necessary, and this translation was made easier by re-framing model metrics to focus on end-users' theory of mind (Section 3.4). Ultimately, we found that the provision of an explicit actionable strategy had a pivotal impact on user strategy when making AI-assisted decisions; it was additionally cited by participants as being the most useful piece of information from onboarding. Thus, model transparency toolkits (e.g. Model Cards [8]) may also benefit from **transforming detailed model metrics into explicit, human-actionable strategies**.

6 CONCLUSION

This case study described the experience of developing onboarding materials for an AI assistant. In addition to providing valuable information to users of the Assistant, the process of producing the onboarding materials served as a means for a cross-functional team to jointly understand how the system would be used, perceived, and assessed in practice. In this capacity, the onboarding materials served as a useful boundary object between all stakeholders—engineers, researchers, and end-users. The information desired in the onboarding materials also affected the development of the underlying model itself, by suggesting improvements to the model in ways most likely to be useful or appreciated by the end-user. These observations suggest the utility of developing onboarding materials early in the design and development process to help align stakeholders and raise awareness of end-user needs and requirements.

REFERENCES

[1] [n.d.]. Cancer of the Prostate - Cancer Stat Facts. https://seer.cancer.gov/statfacts/html/prost.html

[2] Simon Baron-Cohen. 1991. Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading* 1 (1991), 233–251.

[3] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. 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*. 1–12.

[4] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359206

[5] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. 2019. Similar image search for histopathology: SMILY. *NPJ digital medicine* 2, 1 (2019), 1–9.

[6] Bonnie E John, Len Bass, Rick Kazman, and Eugene Chen. 2004. Identifying gaps between HCI, software engineering, and design, and boundary objects to bridge them. In *CHI'04 extended abstracts on Human factors in computing systems*. 1723–1724.

[7] Charlotte P Lee. 2007. Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work (CSCW)* 16, 3 (2007), 307–339.

[8] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[9] Kunal Nagpal, Davis Foote, Yun Liu, Po-Hsuan Cameron Chen, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, et al. 2019. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ digital medicine* 2, 1 (2019), 1–10.

[10] Susan Leigh Star. 1989. The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In *Distributed artificial intelligence*. Elsevier, 37–54.

[11] David F. Steiner, Kunal Nagpal, Rory Sayres, Davis J. Foote, Benjamin D. Wedin, Adam Pearce, Carrie J. Cai, Samantha R. Winter, Matthew Symonds, Liron Yatziv, Andrei Kapishnikov, Trissia Brown, Isabelle Flament-Auvigne, Fraser Tan, Martin C. Stumpe, Pan-Pan Jiang, Yun Liu, Po-Hsuan Cameron Chen, Greg S. Corrado, Michael Terry, and Craig H. Mermel. 2020. (2020). Manuscript under review.

[12] Harold Thimbleby. 1996. Creating User Manuals for Using in Collaborative Design. In *Conference Companion on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada) *(CHI '96)*. Association for Computing Machinery, New York, NY, USA, 279–280. https://doi.org/10.1145/257089.257321

[13] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[14] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[15] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.