"Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human–AI Collaborative Decision-Making

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

Although rapid advances in machine learning have made it increasingly applicable to expert decision-making, the delivery of accurate algorithmic predictions alone is insufficient for effective human–AI collaboration. In this work, we investigate the key types of information medical experts desire when they are first introduced to a diagnostic AI assistant. In a qualitative lab study, we interviewed 21 pathologists before, during, and after being presented deep neural network (DNN) predictions for prostate cancer diagnosis, to learn the types of information that they desired about the AI assistant. Our findings reveal that, far beyond understanding the local, case-specific reasoning behind any model decision, clinicians desired upfront information about basic, global properties of the model, such as its *known strengths and limitations*, its subjective *point-of-view*, and its overall *design objective*—what it's designed to be optimized for. Participants compared these information needs to the collaborative mental models they develop of their medical colleagues when seeking a second opinion: the medical perspectives and standards that those colleagues embody, and the compatibility of those perspectives with their own diagnostic patterns. These findings broaden and enrich discussions surrounding AI transparency for collaborative decision-making, providing a richer understanding of what experts find important in their introduction to AI assistants before integrating them into routine practice.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Human-AI interaction; machine learning; clinical health

ACM Reference Format:
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 (November 2019), 24 pages. https://doi.org/10.1145/3359206

1 INTRODUCTION

Deep neural networks (DNNs) are increasingly being developed for use in medical applications, such as cancer detection and grading [66, 70]). Within this space, a significant portion of recent research has focused on demonstrating that these models can rival the accuracy of medical experts [24]. As these models mature and prove to be reliable and accurate, there is a desire to integrate these capabilities into actual clinical practice. For example, one promising application of machine learning (ML) is its use as a "second set of eyes" to inspect a clinical case [6, 26], with the goal of increasing

Authors' addresses: Carrie J. Cai, cjcai@google.com, Google Research, Brain Team, USA, Mountain View, CA; Samantha Winter, srwinter@google.com, Google Health, USA, Mountain View, CA; David Steiner, davesteiner@google.com, Google Health, USA, Mountain View, CA; Lauren Wilcox, lwilcox@google.com, Google Health, USA, Mountain View, CA; Michael Terry, michaelterry@google.com, Google Research, Brain Team, USA, Cambridge, MA.

2573-0142/2019/11-ART104
https://doi.org/10.1145/3359206

the clinician's accuracy or efficiency. This application of machine learning models falls under the category of a clinical decision support system (CDSS) [27].

Despite the high-performing nature of these algorithms, prior work has shown that CDSSes can be difficult to successfully integrate into practice, citing a lack of HCI consideration as one of the primary reasons for failure [41, 44, 55, 76, 78]. For example, users may resist adopting a tool if they do not understand its capabilities, its intended use, or its utility over existing practices [48, 72, 74]. Algorithmic aversion has also been an underlying challenge for these systems [18, 40, 47]. To address this latter issue, there is growing work aimed at providing users more information behind model decisions at the time of inference [60, 63]. Yet, there are many broader questions that a user may desire to ask of the system as a whole, even prior to observing any specific model decision, such as how well it performs, its potential pitfalls, and implications of use. Additionally, a trained model is the result of numerous design and engineering decisions that users may find useful to know. For example, there are design decisions made with regards to data collection, source of ground truth, and model objectives. These types of *global* transparency questions could be key to forming an accurate initial impression of an ML-based system, and to developing an appropriate mental model to work with it cooperatively.

In this paper, we focus on the initial introductory phase of using a DNN-based diagnostic aid, which we refer to as an *AI Assistant*[1] throughout this paper. Specifically, we investigate the information needs articulated by pathologists during their introduction to an AI Assistant that detects and grades the severity of prostate cancer. While a substantial body of work examines user information needs at prediction time [12, 60, 63], or how to design AI assistance to integrate seamlessly with practices [78], we focus on the initial **human–AI onboarding phase**, when users are first being introduced to an AI system, learning its capabilities, and determining how they will partner with it in practice. Here, we define onboarding as the training necessary for effective human-AI collaboration, to be delivered at the *outset* of a user's introduction to an AI assistant. This information is ideally situated in a larger training program that includes information and tutorials for using the system itself. This initial human–AI onboarding process can be key to initial impression formation and the development of appropriate mental models and strategies of use.

To identify information needs, we conducted a study with 21 pathologists. Each pathologist participated in a semi-structured interview to express what information they desired to know about an AI Assistant before using it. They then interacted with an AI Assistant for cancer diagnosis on digital pathology images, in a think-aloud manner. After performing assisted diagnosis, they continued the semi-structured interview to reflect on the experience and to suggest any additional information that might be useful to know about the Assistant.

Our findings reveal a need for a holistic, global view of the AI Assistant and its capabilities, limitations, and biases, preferably presented in terms relatable to day-to-day practices. Specifically, participants desired information across the following dimensions:

(1) **Capabilities and limitations**: The AI Assistant's overall performance, including its particular strengths and limitations under specific conditions (e.g., well-known edge cases).
(2) **Functionality**: What information the AI Assistant has access to, and how it uses that information to make a prediction.
(3) **Medical point-of-view**: The system's subjective "point-of-view," or its medical style vis-a-vis their own, such as the extent to which it tends to be more liberal or conservative when grading cancer severity.

[1]While there are differing opinions about when and where to use the terms "ML" and "AI," our choice of using the term "AI Assistant" was motivated by the intention of this tool to directly assist a user, as well as the observation that "AI" and "Assistants" are frequently used to describe these types of end-user systems.

(4) **Design objective**: What the Assistant has been optimized for, such as its rate of false positives versus false negatives, and whether it is tuned to compensate for common human errors (as opposed to being as independently accurate as possible).
(5) **Considerations prior to adoption**: The decision factors considered even prior to adopting or purchasing an AI (e.g., effect on legal liability, impact on existing workflows, cost of use).

The breadth and depth of pathologists' information needs indicate that it is not enough to simply provide summary statistics of an AI Assistant's performance (e.g., its accuracy), along with basic training on how to use the interface. Instead, clinicians are likely to relate to the AI Assistant much like they do to a fellow colleague, and ask what its medical point-of-view is, whether it has specific areas of expertise and weaknesses, and how it complements their skill set. Furthermore, there is a clear understanding that any such tool will have its biases and limitations, and that it is critical to learn what these are, prior to use. Collectively, this information not only helps clinicians understand whether they should use the tool, but also how they can most effectively partner with it in practice.

In sum, our contributions are:

- A description of the primary types of information pathologists request in considering how to integrate an AI Assistant with their existing practices (enumerated above and in Table 1).
- Specific examples of how pathologists envision applying that information to collaborate more effectively with diagnostic AI.
- Implications and recommendations for onboarding experts to use AI for collaborative, diagnostic decision-making.

Taken together, these findings provide insight into this initial phase of learning about, and planning to use, an AI Assistant for collaborative decision making, and contribute important implications to the design of the onboarding experience.

2 RELATED WORK

This paper draws on prior work at the intersection of clinical decision support systems and algorithmic transparency. Our investigation of human–AI collaboration is also informed by existing research on collaborative practices in medicine.

2.1 Clinical Decision Support Systems and Computer-Aided Diagnosis

Clinical decision support systems (CDSSes) provide clinicians with knowledge to enhance medical decision-making, such as support for diagnosing patients [27, 54], making prognostic predictions [78], or selecting treatments [80]. In this work, our focus is on computer-aided diagnosis (CAD)[2].

Prior work has identified many challenges of integrating new CAD technology into existing practices. These range from the institutional and sociotechnical barriers, to adoption (e.g. competing stakeholders interests [29] and workflow integration [49]), to the more tactical, in-the-moment challenges that arise during computer-aided diagnosis. Our research is primarily focused on the latter. To this end, a wide range of research has uncovered challenges faced during computer-aided diagnosis [15, 31, 32, 34, 41, 44]. Within pathology, Molin et al.'s design studies found that supporting pathologists' ability to detect errors is vital in this safety-critical environment, and suggest UI design choices for displaying errors in the context of use [52]. In research on early CAD mammography, it was found that users often misjudged the operational scope of these systems [34], assuming them to have a level of interpretive sophistication similar to their own [32]. Users can also

[2]Note that CAD can sometimes refer to computer-aided *detection*. Our use of this abbreviation should be read as computer-aided *diagnosis*, unless otherwise specified.

be surprised by system errors when they were qualitatively different from the types of errors made by medical practitioners [34], or when they are inconsistent with their own mental schema [33]. Collectively, this body of work suggests a need for properly setting expectations and prescribing appropriate use of these computational tools.

While the majority of this prior research examined how experts *use* CAD systems, our work specifically focuses on eliciting how clinicians would want to be *onboarded* to such systems: our study, for example, intentionally elicits onboarding desires prior to (and at the very outset of) actual tool use. Most relevant to our work is early research on mammography prompting systems, which found that reporting error modes as part of training helped users better understand prompt errors, but providing best practices did not prevent them from misinterpreting the intended use of the tool [32]. Building on this early work, our research provides a much deeper examination of onboarding, at a critical juncture in the history of CAD: whereas first-generational CAD systems focused primarily on *detecting* suspicious regions (e.g. in Hartswood et al.'s aforementioned work), the rise of modern-day machine learning has enabled new CAD systems that not only detect, but also *interpret* and *diagnose* those regions. This arrival of high-performing, deep learning-based diagnostic systems, coupled with medical experts' increasing familiarity with CAD systems, suggests a pressing need to re-examine how best to introduce medical experts to these ML-based tools: not only are the diagnostic nature of AIs likely to create new questions from end-users, but the attitudes that practitioners have now developed toward early CAD systems may also frame and shape new onboarding needs. As such, our work contributes an in-depth investigation of onboarding for ML-based diagnostic AI assistants. In doing so, it also provides a clinical angle to recent human–AI interaction guidelines, which state that AI systems should make their capabilities clear to users upon initial interaction with the system [5].

2.2 Algorithmic Transparency

The growing prevalence of deep learning models and their use in high-stakes decision making have led to increasing demands for more transparent and explainable AI. However, the interpretation of deep learning models is challenging due to their complexity and often opaque internal state. To address this, the machine learning community has produced a myriad of algorithmic and mathematical methods to explain their inner workings. These methods aim to explain the model prediction outcome for a single input data point [21, 43, 63, 67], or for a set of data points in a predicted class [42], often by perturbing model inputs and seeing how the model's response changes. Across this work, there is a clear need to ensure usability and efficacy with real users [1]. In light of this, a recent wave of HCI research has studied what end-users actually desire to understand about ML systems, and how that transparency affects user attitudes and outcomes. Domains of study include recommender systems [4], medicine [12, 75], social media [22], creativity [14], and advertisements [23]. While the majority of this work has tended to focus on explaining the reasoning behind specific model decisions, our work instead examines the broader questions that a user may desire to ask of the system as a whole, including components of the ML pipeline that may occur even before a model is built (e.g., data collection, or selection of model design goals).

Recent work on fairness and accountability has proposed the use of short documents accompanying trained ML models to disclose the intended use of models, details of their performance evaluation procedures, and the potential biases that they may embody [25, 51]. Relatedly, others have recently studied whether laypeople's trust in a model varies depending on the model's stated accuracy and how it differs from observed accuracy in practice [79]. Our work builds on this growing interest in studying the broader, global aspects of model transparency, and conducts a deep exploration of these issues within the domain of medical decision-making.

2.3 Collaborative Work in Medicine

The design and study of collaborative work in health care has been widely researched. Prior work has examined the shared *awareness* maintained between medical experts during collaborative work [8, 58, 61, 62, 65, 71]. Bardram and Hansen observed that mutual awareness of the activities and whereabouts of other clinicians is central to determining when and where to seek advice from colleagues [7]. Work on *articulation* and coordination of work components [2, 13, 50, 81] is also extensive. Within this space, Larsen and Bardram found that *competence articulation*, the articulation of one's competence to a co-worker, improves collaboration during telemedical consultations by allowing clinicians to utilize each others' expertise [45]; such findings are in line with foundational social psychology theories [17, 77], which describe how collaborators form and leverage mental models of each others' capabilities. Within medical imaging, there is evidence that the informal sharing of expertise is fundamental to collaboration. Examples of this include practices of developing a familiarity with others' awareness of local conditions [37], or making one's work visible to others when multiple people need to examine a single scan [30]. Finally, research on the clinical *appropriation* of interactive technologies into working practice found that technical systems need to be amenable to internal examination and reasoning by end-users, in order to be incorporated in their work [19]. Building on this large body of work, we examine what types of information end-users desire to know about an AI Assistant during onboarding, and relate these needs to the existing medical practices of competence articulation and the seeking of input and second opinions from colleagues.

Prior work also examined the specific challenges of integrating CAD into diagnostic practices that are inherently collaborative. For example, Hartswood et al [31] found that existing artifacts and practices (e.g. protocols and annotations) facilitate socialization between clinicians, and that CAD technology can be challenging to integrate if it lacks this ability to reason through decisions or to be queried. Jirotka et al. [37] examined distributed digital mammography systems and found that, because readers orient to local practices, they needed an ongoing understanding of each others' familiarity with those local processes; likewise, CAD technology may need to express the extent to which it aligns with local standards. At the organizational level, integrating new technology requires overcoming the hurdles of existing in professional cultures and institutional politics (e.g. competing interests by different stakeholders) [29]. In sum, these collaborative diagnostic practices form a basis for understanding what users may need to know about an AI during onboarding to develop a productive working relationship with it, and what challenges may need to be overcome given existing dynamics in human collaborations.

3 BACKGROUND

In this paper, we studied the onboarding needs of pathologists, who diagnose diseases (e.g., cancer) through the microscopic examination of tissue samples. While human–AI onboarding is relevant to many medical fields, the application of machine learning to pathology could be particularly impactful: whereas clinicians may request specialty referrals or additional tests to provide more evidence if they are uncertain of a diagnosis, pathologists are often responsible for providing a final, tissue-based diagnosis, especially for cancer. Here, our work is focused on a prostate cancer use case. Prostate cancer is one of the most common types of cancer, and the subject of significant attention in recent work on machine learning-powered cancer diagnosis [11, 15, 35, 56, 57, 69].

Pathologists diagnose the severity of prostate cancer through the Gleason scoring system [20]. The Gleason score is one of the most important predictors of prognosis, and is widely-used clinically to guide patient management decisions. However, like many aspects of medicine, Gleason scoring also involves an unavoidable degree of subjectivity, and can suffer from suboptimal interobserver

variability [59, 73]. Because different disease severity levels can adopt similar visual patterns and features, the same visual patterns can be interpreted differently by different pathologists, due to how they apply standard guidelines. Thus, even though there is arguably an objective truth (whether cancer exists or not), variability in cancer grading exists due to differences in the interpretation and application of guidelines to a given case; for the time being, a subjective element to cancer diagnosis is inevitable.

Inherent to nearly any diagnostic process is a process of collaboration, both across disciplines and among specialists of the same discipline, often with the goal of reaching a consensus decision [38]. For breast cancer, the multidisciplinary process has even been formalized into a well established "triple test" scoring system [39, 53] to integrate information across multiple diagnostic modalities (clinical findings, radiology, pathology). For prostate cancer pathology, the histopathologic diagnosis typically stands alone more independently as a diagnostic "ground truth," a notion which likely results in important differences regarding the overall collaborative nature of the diagnostic process and user interaction with potential CAD systems. While staging and treatment decisions do integrate clinical history, laboratory tests, and radiological imaging, these other modalities are often less reliable for the actual prostate cancer detection and grading, such that the diagnosis is typically rendered specifically based on the histopathological specimen. In this setting, the most common collaborative behavior when uncertain about a case is to seek additional input through consultation with another pathologist, either an immediate colleague or sometimes from an outside institution or an expert pathologist specializing in prostate cancer, known as a Genitourinary (GU) pathologist (also referred to as urological pathologist). Taken together, the existing collaborative practices of medical diagnosis, combined with a diagnostic reliance on the pathologic interpretation and subjective nature of Gleason scoring, introduce unique challenges surrounding human-AI collaboration and decision-making that are worthy of deeper investigation.

4 METHOD

The driving research question for this work was to document what pathologists want to know about an AI Assistant prior to its use. In large part, our goal was to inform the design of onboarding materials for AI-based assistance in pathology. To this end, we were also interested in understanding what existing mental models participants held for AI Assistants (e.g., to ensure any onboarding materials appropriately address potential misconceptions or gaps in knowledge), and to understand how working with an AI Assistant may be similar to or different from working with colleagues or existing tests and instruments.

To address these research questions, we first conducted open-ended interviews with three pathologists to inform a set of semi-structured interview questions. Then, we conducted a three-phase qualitative laboratory study with 21 pathologists, distinct from those who participated in the initial interviews. In the first phase (*pre-probe*), we asked participants what types of information they would need to know about an AI Assistant before using it. We also asked them to describe how they were previously onboarded to an existing technology or diagnostic test from their current practice. In the second phase (*probe*), we sought to understand users' information needs while making a decision with an AI Assistant. Participants diagnosed a prostate case with the aid of an AI Assistant: in line with sequential read procedure [64], they reviewed the case immediately before the algorithmic predictions were revealed. The algorithmic predictions (benign, grade 3, grade 4, and grade 5) were displayed as visual overlays on the image (Figure 1), which the pathologists could consider when making a final diagnostic decision. Images were from the TCGA Research Network's pathology image repository[3]. In the third phase (*post-probe*), we asked participants what additional

[3]https://tcga-data.nci.nih.gov/docs/publications/tcga

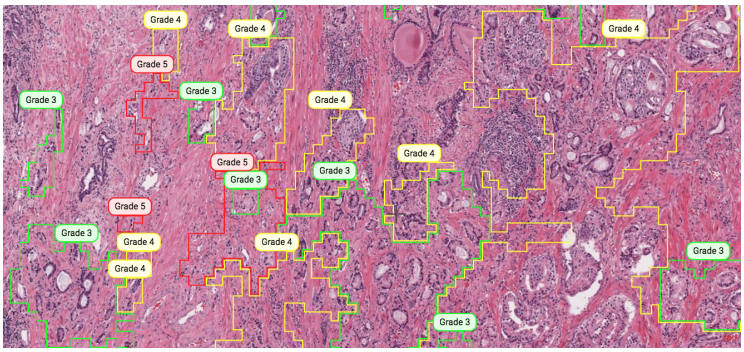

Fig. 1. As a study probe, we presented participants with prostate cancer grade predictions (e.g. grade 3, grade 4, and grade 5) made by a high-performing deep neural network. The predictions were shown as colored overlays on top of the prostate tissue. The predictions are made at the region-level because, in regular practice, pathologists identify region-level grades and patterns to produce a final, overall Gleason score.

information they felt they needed to know about the AI in order to work with it effectively. Finally, we asked how they envisioned being onboarded to using such an AI Assistant. Each session lasted between 1-1.5 hours.

In identifying a prostate cancer case to use as a probe, our goals were two-fold: the case should be non-trivial to diagnose, and the model's behavior on the case should also be representative. To ensure the task was non-trivial, we first identified cases that had previously been contested between pathologists, with conflicting diagnosis labels. Because some grade differences have minimal impact on clinical treatment, we further filtered for cases where the differences in opinion represented clear differences in clinical impact. In pilot studies, and through conversations with developers of the AI model, we identified the most common peculiarities of the algorithm (e.g. difficulty on regions containing processing artifacts), and ensured that the case probe captured these common model behaviors.

All interviews were screen recorded and transcribed. Five researchers independently familiarized themselves with the data, and then met as a group to collaboratively generate an initial set of codes in a bottom-up fashion. Then, two of the researchers revisited the data and identified important sections of text to attach to codes, working systematically through the data set to apply and refine codes in a collaborative, iterative process. The five researchers then met again to re-examine codes, which formed the basis for themes across the data set. The researchers examined relationships between codes, and iteratively converged on a set of themes [9, 10].

4.1 Participants

21 pathologists participated in the study. All were general pathologists recruited from among a pool of remote contractors assisting with pathology projects at our institution. Their main practices spanned a wide range of settings and institutions, including government-funded / community hospitals (7), private practice (6), academic hospitals (3), private hospitals (2), consulting (2), and independent laboratories (1). The size of their institutions also ranged widely: the reported number of general pathologists on staff ranged from 1 to 120 (mean=13). Most of their institutions (14) did not have any GU pathologists on staff, and the rest had between 1 and 8. Most participants reported that they sign out prostate cases several times a week (13), one reported several times a day, and the rest encountered prostate cases several times a month or less (7). This is in line with expectations,

given that general pathologists tend to see all types of pathology cases, not limited to prostate. Overall, participants had 1-25 years (mean=10.8) of pathology experience post-residency training.

4.2 Limitations

Our study provides an in-depth investigation of initial onboarding needs, but it does not capture long-term effects or how these needs might change over time. Furthermore, because participants were recruited through our institution, and were peripherally involved with pathology data-collection projects, they may already be aware that machine learning is to some extent based on data. Since this may narrow the types of information participants ask for, our findings should be interpreted as representing a lower bound on the scope of all onboarding information needs. As described above, their main institutions – where they spend the vast majority of their time – were diverse in both size and clinical setting. Moreover, because we used a single case probe, there may be other model behaviors or onboarding needs that were not captured as a result. We held consistent the probe to ensure protocol consistency between participants, and ensured that the case captured the most commonly occurring characteristics of the model. As shown above and in the rest of the paper, the information needs reported were quite broad and rich, even in light of this limitation.

To investigate what information pathologists felt they needed to effectively partner with an AI Assistant, our study design included think-alouds to observe the diagnostic process, and questions to help paint a picture of the larger diagnostic process and the people, tools, and resources they interact with during that process (e.g., how they currently obtain a second opinion). While this study design provided a window into their diagnostic practices, one limitation of this design is that we did not observe diagnosis in situ. As a result, aspects of their collaborative processes (actual collaborations with colleagues, interactions with other tools, artifacts, and resources) may be under-reported or absent from the data we collected. These data could provide additional insights into the types of information and activities that pathologists might find useful when learning how to effectively collaborate with an AI Assistant.

5 TYPES OF INFORMATION DESIRED AND HOW THEY INFORM COLLABORATION WITH AI

As presented in the Introduction, we frame our findings around the five themes that emerged from participants' responses about their information needs: capabilities and limitations, functionality, medical point-of-view, design objective, and considerations prior to adoption (Table 1). After presenting these information needs, we then describe how participants envision applying this knowledge to reconfiguring their work with an Assistant. Table 1 summarizes our findings.

Many of the issues brought up by participants are relevant to any tool or instrument introduced into a medical context. For example, there will be obvious and necessary concerns surrounding the system's accuracy across representative cases. However, other concerns are arguably more specific to an AI Assistant, such as the medical point-of-view represented by the AI Assistant (e.g., due to choices in training data or labeling). For completeness, we report all of the information participants desired to know about an AI Assistant, but highlight information particularly germane and unique to AI Assistants.

5.1 Capabilities and Limitations

For effective collaboration with AI, participants described a wide range of information needs surrounding desired knowledge of its performance and limitations.

5.1.1 Accuracy and Human-Relative Measures of Performance. Accuracy of the AI Assistant was a primary concern of participants, with some stating a minimum level of accuracy expected (e.g.,

Table 1. Summary of findings and implications for onboarding.

Capabilities and Limitations
• Anchor performance metrics to human-relatable benchmarks • Define accuracy precisely given multiple interpretations (e.g. binary benign vs. cancer, multi-class accuracy, percent tumor, etc.) • Gather known human pitfalls (e.g. well-known edge cases) and report AI performance on those sub-categories • Describe the diversity (or lack thereof) of the training data to inform generalizability • Relate volume of training data to what is considered reasonable scale for machine learning • Describe theoretical limits of AI, given current knowledge
Functionality
• Enumerate the inputs / context accessible to the algorithm, particularly inputs that are not shown in the interface (e.g. patient history) • Specify the main steps in the AI's analysis of its inputs, vis-a-vis steps taken in typical human analysis (e.g. multiple magnification levels, decomposition of input into sub-images) • Compare and contrast AI schemas relative to known human decision-making schemas (e.g. the extent to which it has been explicitly trained on higher-level biological concepts)
Medical Point-of-View
• Show subjective thresholds of the model (e.g. examples of AI judgment on borderline cases) • Include a human-AI calibration phase, for users to calibrate their own subjective thresholds to that of the AI, with an authoritative source provided as ground truth • Specify where the algorithm received its medical source of ground truth (e.g. expertise level and number of clinicians, prognostic data, etc.)
Design Objective
• Make explicit the AI's intended utility over the user's status quo (e.g. efficiency, accuracy, consistency, etc.) • Make transparent whether the AI accounts for unequal cost of errors in its objective function (e.g. false positives vs. false negatives; differential impact on treatment) • Indicate whether the AI is explicitly tuned to work with a human partner (e.g. designed to complement certain human weaknesses), versus independently optimized
Considerations Prior to Adoption
• Provide information on regulatory approval (FDA), peer-reviewed publications validating the tool, impact on existing clinician workflows, impact on legal liability, and cost of purchase
How the information above could inform collaboration with AI assistants
• **Attention allocation**: User dedicates more attention to AI areas of weakness, and less energy to AI areas of strength • **Conflict resolution**: When user and AI opinions conflict, user considers what context the AI has access to, and how it processes that information, when determining how much to trust the AI versus themself • **Upgrading and downgrading**: User takes the AI's subjectivity and biases into account when considering its suggestions to up-grade or down-grade disease severity • **Mode of collaboration**: Knowing what the AI is optimized for (e.g. complementing human weaknesses at the cost of raw accuracy) sets user expectations for the mode of collaboration

95%). Although participants were told that the Assistant predicts Gleason grades, many assumed that accuracy referred to the binary classification of benign versus cancer, suggesting a potential unfamiliarity with the subtleties of assessing the performance of multi-class classification systems. This finding suggests that it may be useful to determine the kinds of performance metrics users are accustomed to seeing, and highlight any differences in the definition of accuracy being used in onboarding so that stakeholders and end-users are better equipped to understand empirical measures of an AI Assistant's performance.

While participants naturally insisted that the AI Assistant be accurate, many were not sure what should constitute a reasonable performance threshold. Instead, they desired to contextualize and compare its behavior relative to "human benchmarks." For example, participants wanted to know how its diagnoses correlate with a panel of GU pathologists; what its error rates are relative to personally observed human error rates; or how it compares to published rates of concordance between general pathologists. In sum, pathologists desired empirical measures of performance, but in order for these metrics to be meaningful, they may need to be reported relative to human benchmarks, and with a precise definition of metrics used (e.g., how accuracy is being defined).

5.1.2 Common Pitfalls and Known Edge Cases. Whereas initial needs tended to center on high-level performance metrics, participants eventually expressed a deeper desire to understand the AI's specific categorical strengths and limitations. The most common desire was to understand the pitfalls of the AI system, so that they could anticipate those weaknesses and account for them during decision-making: *"What is difficult for the AI to know? Where is it too sensitive? What criteria is it good at recognizing or not good at recognizing?"* (P18) Participants described parallels to the current onboarding materials they use in medical practice, which typically highlight known pitfalls and limitations: *"There's a prep for pap smear and...at the edge, the cells dry out and look bigger, so it's a known thing to not evaluate or you might think there's dysplasia where there's not. So they point that out in the training materials."* (P4) Others described learning about pitfalls from colleagues who have time-tested experience using the technology.

Many participants were able to describe specific scenarios that the AI Assistant should be validated against. These scenarios are known to be difficult for humans, such as benign mimickers of cancer (e.g., atrophy) or special sub-patterns (e.g., cribriform, perineural invasion): *"Maybe it has really good accuracy except for perineural invasion. If you see perineural invasion...don't fall for that."* (P20) Instinctively, pathologists often assumed that the AI would have difficulty with the same special cases that they themselves struggle with. However, they usually gave themselves more credit in being able to properly handle these cases, referring to their own perceived ability to "correct" for those exceptions: *"I would call it not-interpretable. Most pathologists aren't going to interpret something right at the tissue margin."* (P17) Several participants eventually wondered whether the AI could have already corrected for those factors as well: *"It's important to know that the AI is correcting for it or to know that it's a flaw."* (P4). Given these well-known special cases, one could imagine stating in onboarding whether the AI has been trained to handle known edge-cases, to aid participants in building correct mental models. We elaborate on this possibility in the Discussion section.

5.1.3 Training Data and Generalizability. To understand the AI Assistant's likely capabilities and limitations, some participants desired a summary of the volume and types of clinical cases that the algorithm was created from. However, participants felt they did not have a benchmark for evaluating what volume of data would be adequate, or what scale is reasonable for machine learning: *"I am not an AI expert, so I cannot point out the number."* (P16) Some suggested that the number of data points should be on par with the volume of cases pathologists are typically trained on, with

some feeling that it would need at least enough cases to have observed the *"rare variants that we come across once every 4 or 5 years."* (P13)

The desire for high case volume may reflect a desire for generalizability. For example, one pathologist pointed out that having data from diverse sources would be more representative: *"More variation is better...Covering from community hospital small groups, to academic medical centers, it's more representative."* (P16) Another wondered if an AI that is *"geared toward a certain type of stain"* could generalize to stains at other institutions: *"Our staining is really bad and there are days when the stain is faded...In those circumstances, and with different variables, how good is that system going to be?"* (P13) Overall, providing users a sense of the diversity of training data could help inform generalizability. However, numerical metrics about case volume may need to be accompanied by benchmarks to give users a basis for what is reasonable within the scale of machine learning.

5.1.4 AI-Specific Capabilities and Limitations. As they discussed different means of assessing the capabilities of an AI Assistant, participants also expressed a desire to obtain a basic understanding of the ultimate limits and capabilities of AI in general. As an example, one pathologist described the human ability to have an inkling or "sixth sense" that cannot be rationalized, and desired to know if AI could theoretically ever capture such a human instinct: *"Are there certain things that are natural limits to the technology that could never be supplanted by AI? Or is it just my human egotism? Is the sixth sense an illusion? ... Maybe I'm romanticizing it or I'm being delusional."* (P18) Others reasoned that the AI couldn't possibly capture elements that are imperceptible to the human eye, such as proteins that currently can only be detected through staining techniques. Still, some maintained that there are no limits to what an AI can learn, so long as there is sufficient data to learn from: *"As long as they learn enough and can correct itself enough, with enough data, then it can be perfect."* (P16) Although the theoretical bounds of AI can be challenging to identify given that the field is still evolving, onboarding materials could at least offer the current state of knowledge, to set realistic expectations.

In sum, participants desired to know not only the model's aggregate performance, but also its specific categorical limitations and potential pitfalls, especially on well-known edge cases that are difficult for humans. Given variation in performance expectations and lack of knowledge about AI, numerical metrics about model performance or training data should also be framed in relation to human-understandable benchmarks.

5.2 Functionality

Differential diagnosis is a process that combines several information streams (e.g., tissue samples, patient history, lab results) to arrive at a final decision. In this section, we describe what pathologists desired to know about system inputs, how the AI processes and analyzes those inputs, and how the AI arrives at a decision.

5.2.1 Inputs and Accessible Context. When using the AI Assistant, participants naturally inquired how much context is available to the Assistant. For example, one participant wondered whether it had analyzed other data from the same patient, such as patient history or additional tests: *"Does the AI assistant have access to information that I don't have? Does it have access to any ancillary studies?"* (P10) One common question was whether the input available to the AI mirrored exactly what was shown in the user interface (a single image), which contains less context than the multiple levels of images pathologists typically consider when examining a case: *"I want to know if the AI is being generated off of one image or if it's being generated based on sequential images – the levels. Sequential I would trust more."* (P14). Without the ability to see the "bigger picture," participants felt it was inherently limited in its abilities: *"It's unfair for AI because it just does whatever it was set up to do, so it doesn't get a chance to get an overall thing of the big picture."* (P14) Given these insights, it

would be important to make explicit to users what input the AI does and does not have access to, particularly if it takes in more information than what is immediately shown in the interface.

5.2.2 AI Analysis of Inputs. In clinical practice, a pathologist examines a slide at multiple levels of magnification. Accordingly, participants sought to understand how the image was examined by the system, and whether it matched their own practices: *"[Is it] looking mainly on the low power view, or integrating everything? Is it scanning low power and high power?"* (P18) Because some might assume that these nuances of practice are only learned through human medical training, it could be important to state upfront how the AI assesses its input, particularly in relation to existing human clinical practices.

In other cases, a lack of knowledge about how the AI processes inputs could lead to a rapid degradation of trust. For example, this particular implementation of the AI Assistant first segmented the image into smaller image patches, then performed prediction on each patch. Consequently, a well-defined biological structure could occasionally be arbitrarily segmented into two regions and have two different predictions assigned to it. Some participants were bewildered by this behavior: *"Why did it cut this gland? It should circle the whole gland, not just part of the gland...This does not exist in nature."* (P16) Not knowing the AI's process degraded trust, even if the AI's predictions would have otherwise been valuable had the clinician been able to look past this behavior.

5.2.3 The AI's Decision-Making Process. To arrive at its decision, many participants imagined the AI employed a nearest-neighbor lookup: *"I think it just uses the picture and compares with what they have in their database and sees which one it fits. Find something that looks similar in this database."* (P14). Some postulated that, because the computer classifies images into discrete categories, it is not aware of the continuity between cancer grades: *"Pathologists know these patterns can blend, but the computer's doing the best to fit into a 4 or a 3."* (P12) In understanding how it arrives at its decisions, opinions were mixed as to what concepts the AI actually "knows" through its training. For example, some guessed it could only learn visual patterns derived from basic visual elements (*"Maybe light and dark? Maybe colors? Maybe shapes, lines?"* (P17)), whereas others wondered if it could learn higher-level biological concepts (*"Does it take into consideration the relationship between gland and stroma? Nuclear relationship?"* (P16)). Without a sense of what level of abstraction AI is capable of learning, pathologists struggled to determine the extent to which its diagnostic process could be similar to or different from their own.

This lack of an understanding of how the system arrives at a decision led to the desire for an AI primer: *"If someone can explain in a simple language, this is how it does it...so that intellectually we can understand what's going on in [the] AI's brain, and compare to our brain."* (P1) Participants contrasted this tool to other clinical technologies for which they typically already have a conceptual foundation through years of residency and training: *"When I bring on a test, I usually know what method it is. You tell me AI, and I have conceptually no idea."* (P17) As a result, pathologists wanted to get a basic crash course in using AI, with some even acknowledging that such a course would be an essential prerequisite to practicing modern medicine: *"Generationally, it's going to be one of the more important ideas that's happening. How it works, just a primer."* (P18)

In sum, pathologists desired to know the inputs and context accessible to the algorithm, as well as the basics of how it analyzes those inputs to make a decision. Their natural inclination to consider AI behavior in relation to their own human schemas suggests that an AI primer could be particularly fruitful if it compares and contrasts AI processes to that of known human clinical processes.

5.3 Medical Point-of-View

In addition to questions and theories regarding how the AI functions, participants also had numerous questions about the medical point-of-view embodied by the Assistant. They noted the subjectivity inherent in cancer grading, motivating their need to understand both where the algorithm lies on this spectrum of subjectivity, as well as where the algorithm receives its source of ground-truth.

5.3.1 Subjectivity in Current Clinical Practice and Implications for AI. Many participants noted the subjectivity intrinsic in assigning cancer grade levels: *"There's a lot of subjectivity, grade 4 vs grade 5, you ask ten pathologists, four will say one thing and the rest will say another."* (P18) As disagreements are common and variability is well known, pathologists typically rely on their knowledge of each others' grading tendencies when seeking a second opinion: *"I know one of my colleagues will call almost anything high grade, and the other one does not like calling high grade at all."* (P7) Knowledge of these tendencies helps them decide who might be appropriate to ask for a second opinion. Often, they sought out colleagues with grading tendencies similar to their own, so that when a disagreement arises, they can trust the legitimacy of those disagreements: *"If the person you do normally see eye-to-eye with tells you that it's wrong, then you'll believe that more."* (P20) To many, an ideal collaborator would be one who shares similar medical points of view as oneself, but who also adds additional insights: *"What are they worried about? And I'll think, was I worried about those areas as well?...If they overlap [with mine] but then add a little bit [of extra information] ...I think that would make me more trusting."* (P15)

This practice of learning the clinical styles of one's colleagues led pathologists to inquire about where the algorithm lies on the spectrum of subjectivity, similar to how they might calibrate to their peers: *"It'd be interesting to know how you calibrate your eye to a system that you're going to be using. What are you calling and what is it calling? I know what my friend...will call, and what I call...What would AI call it?...I'm treating it as a peer."* (P18)

As with choosing peers for second opinions, participants expressed a desire for the algorithm to have similar diagnostic styles to oneself, such as being more liberal or more conservative in assigning higher-severity cancer grades: *"Does it have a bias a certain way? I kind of want it to think the way you do...I certainly wouldn't want it to be completely discordant with what I think are the subtle nuances between 3 and 4."* (P2) Having similar subjective thresholds could be key to developing trust, so that they could later rely on it when uncertain: *"If I find every time I call [grade] 4, it calls 4, then like a person, you build up trust in each other. Then next time if I can't decide, then I would trust the computer. That would build up trust, it's very important. Let's say I use it for one week, and 30% of the time I don't agree, then I cannot trust 4 is 4."* This suggests that the subjective operating points of a model may need to be made transparent or even adjustable by end-users. However, there may exist a tension between the subjective alignment with users needed to establish trust, and the shift in current user biases needed to improve clinical practice.

5.3.2 Medical Background of the AI Assistant. Due to the subjectivity inherent in cancer grading, participants desired to know from whom or where the algorithm received its medical source of ground truth. Whereas pathologists are typically aware of their colleagues' medical experience and pedigrees, the clinical background of an AI can be opaque. Participants asked whether the algorithm had learned from diagnoses made by general pathologists, GU pathologists, or an entire panel of GU pathologists. Most felt they would not trust an algorithm unless it were based on judgments made by well-respected GU pathologists or institutions, explaining that expert consultation is how they typically resolve uncertainty in current practice: *"I would send it to [name]. He has 50 yrs of experience in GU pathology and wrote [a critical text in pathology]. So you're getting someone with a higher level of expertise to weigh in."* (P7) A few participants asked if the AI was based on an even

more objective source of truth than GU pathologists, such as patient prognosis or immunostains. Overall, due to variability in grading, pathologists may not know how to assess the qualifications of an AI unless it is known to be based on an authoritative source of truth.

5.3.3 Human–AI Calibration. To develop an understanding of the AI's grading competence and potential biases, some pathologists envisioned assembling a set of cases with ground truth (e.g., as assigned by GU pathologists), and comparing their diagnoses and the AI's diagnoses with the ground truth in a calibration phase. This practice would serve dual purposes: It would give insight into the AI's diagnostic tendencies, as well as their own: *"[The] AI would look at a slide, I would look at a slide, and then we would know what the expert said. I think that would be interesting just to see where I stand."* (P17) Pathologists felt this feedback would be valuable particularly because they currently lack formal, personalized feedback on subjective thresholds in current practice; they explained that standardized exams tend to only test classic scenarios rather than grey area cases, and that most feedback is received through inductive observation when getting a second opinion.

Several participants cautioned that any such human–AI calibration sessions should be tempered to be as non-confrontational as possible: *"The average pathologist, if you say, 'Oh you're wrong,' they're going to be uncomfortable with that."* (P14). Rather than providing a direct comparison between the clinician and the AI's performance, they recommended more implicit approaches that make the calibration feel less like a comparison, and more congenial: *"When it does better than you, it's always nice to have a gentle safe environment...Hopefully it's done in a way where it's positive, constructive."* (P2) Participants suggested that, rather than providing a raw numerical comparison, it would be much more meaningful and actionable to illustrate *what* those differences were, as case examples.

In sum, participants desired knowledge of the AI's subjective tendencies, much like how they might consider the diagnostic patterns of their peers when seeking a second opinion. Many felt a human–AI calibration phase would be valuable, if done in a positive and non-comparative way. Critically, they also needed to know the algorithm's source of ground truth in order to trust its authority. A calibration session during onboarding could allow users to "practice with" the AI assistant to develop an understanding of the AI's subjective stance and clinical perspective.

5.4 Design Objective

Pathologists also recognized that the AI Assistant will have a particular medical point-of-view due to its designed, intended objective. Although all participants were informed that the Assistant predicts prostate cancer grades, almost all desired an explicit statement on the specific *utility* of the tool—in other words, how the AI was intended to benefit them over their existing practices. Expectations for the intended utility ranged from increasing efficiency and accuracy, to ensuring consistency and reducing fatigue.

Noting the possibilities, multiple participants debated the inherent trade-offs that the designers of intelligent systems must navigate in implementing the system. For example, some brought up the example of existing automated screening tests (e.g., pap smear), which are often optimized for greater *sensitivity* at the cost of decreased *specificity*, given that positively flagged cases are manually reviewed by physicians. Aware of the different decisions that could be made, some participants asked if the Assistant had been optimized for a particular metric: *"The most important thing is just seeing how it was tuned..., just the ROC...whether it's meant to make some trade-offs in terms of how sensitive it is. It would be nice if we knew it was very sensitive, wasn't going to miss things...might have some false positives but you're going to look at that."* (P2) In addition to sensitivity versus specificity trade-offs, some asked if the algorithm had been tuned to consider the differential impact of different grading errors on patient care, since each cancer grade cut-off suggests different

treatment plans. While a growing body of HCI work argues strongly for encoding such human values into AI technology [46, 82], our findings suggest that those algorithmic design decisions should be made transparent to users, along with potential implications for those decisions.

Interestingly, a few participants wondered if the algorithm had been tuned to account for the fact that it would be collaborating with a human end-user, as opposed to operating on its own. For example, when the algorithm's demarcation failed to fully circumscribe a region of cancer, including only part of it, one user hypothesized that, *"Maybe it's just trying to draw my attention to the areas of cancer, and I'm still supposed to decide."* (P20) These discussions suggest that end-users may desire information about an AI's *theory of mind*—its awareness of another being's knowledge and intents [16]—when making a diagnostic decision. For example, some wondered if *"it was thinking about how I would look at something."* (P18) Rather than viewing the AI as an independent decision-maker, these participants instead considered what the AI may or may not know about their own modus operandi.

In sum, it may be useful to not only provide an explicit statement about the AI's intended utility, but also make transparent whether the AI is tuned to optimize for certain objectives, whether it considers the differential costs of errors (e.g. differential impact on treatment), and whether it is explicitly designed to complement a human partner.

5.5 Considerations Prior to Adoption

Beyond the information needed to make effective use of AI, pathologists also brought up several key factors that would influence their initial decision to adopt or purchase such a tool to begin with. These include: evidence of FDA approval and published validation in peer-review journals, social endorsement by well-respected medical leaders, impact on existing workflows, impact on legal liability, and cost of purchase. While the focus of this paper is on the onboarding needed for effective use of AI, rather than the initial decision to purchase or adopt it, we view these initial pre-requisites to be critical to adoption, and should thus be included in onboarding as well.

5.6 Re-envisioning Work with the AI Assistant

The information pathologists desire enables them to consider whether they would adopt the tool, and if adopted, how it could be applied in practice. In this section, we describe how they imagined this information about the AI Assistant would inform their use of the tool and potentially alter their work practices.

5.6.1 Attention Allocation. Participants described how an awareness of the AI's strengths and weaknesses could support their strategic allocation of attention. Some compared this to the ways in which they currently allocate energy depending on the known expertise and weaknesses of their coworkers: *"It'd be like working with a partner. Like I know what my coworkers are strong and weak in...I probably anticipate weaknesses and pitfalls and dedicate more mental energy towards trying to fill in the gaps...I'd develop a working relationship with AI where my awareness is heightened knowing the AI's weaknesses and dampened with the AI's strengths. I'd develop a symbiosis with it."* (P18) Several described how they would dedicate less energy toward scenarios where the AI is known to do well (*"If I employed AI, I would probably just focus on a few areas here."* (P18)), and dedicate more attention and care toward areas that are known weaknesses of the algorithm (*"Things to watch out for that it might miss ...you should actually still put it at higher power and look around for yourself."* (P20)). These descriptions support prior work on competence articulation in medical teams [45], which enable colleagues to leverage each others' expertise and complement each other, thereby enhancing collaborative work.

5.6.2 Conflict Resolution. Pathologists also described how they would use their knowledge of the AI to resolve situations in which their own opinion conflicts with the AI's predictions. This need for disambiguation is critical to human–AI decision-making: when a pathologist's opinion differs from that of the AI, the pathologist must determine the extent to which to trust the AI's opinion over their own, while operating under uncertainty: *"It's hard for me to know how much to trust when I see things that I don't completely agree with."* (P10) In these scenarios, pathologists described how they would use their knowledge of the AI to come to a conclusion. For example, one pathologist explained how knowing an AI's over-sensitivity to benign mimickers could help them discount its opinion in those situations (*"I would earmark that...Say it said 'look at this', I would say, oh it would typically flag atrophy which is benign."* (P18)). Conversely, knowing that the AI had additional context could increase reliance on its opinion, in cases where that context matters (*"If it looked at immunostains, then it has more information than I do. That would make me trust it more."*). Overall, understanding the AI's strengths, limitations, and functionality could be critical to reaching a decision resolution when their opinion differs from that of the AI.

5.6.3 Up-grading and Down-grading. During the study, pathologists also described how an awareness of the AI's subjective decision thresholds could help them determine how much to trust its judgments, similar to how they currently calibrate to their colleagues' tendencies: *"I would have a degree of understanding of the degree of swaying that it's doing...Like if it calls 4, but I know, oh yeah, but this calls 4 on everything, then I can kind of dial it back in my mind...Just like another staff or resident, like this guy always calls 4, this guy never calls it. I would have a mental image in my head of what kind of pathologist is this Assistant."* (P15) Because the distinction between diagnostic grades is often continuous rather than discrete, having the ability to calibrate to an AI's subjective idiosyncrasies could be critical to its effective use. Some felt that a human–AI calibration phase during onboarding, with expert judgments provided as ground-truth, could help them know when to trust the AI's subjective thresholds over their own and even adjust their own thresholds over time: *"I guess I would have more faith in the AI...if maybe my concordance with experts on pattern 4 is low, but its concordance with experts on pattern 4 is high...then I could maybe adjust my thresholds for calling something 3 versus 4."* (P20)

5.6.4 Mode of Collaboration. Finally, transparency around an algorithm's design objective could affect the user's expected mode of collaboration. During the study, the algorithm's predictions were occasionally over-sensitive, marking a cancerous region along with some surrounding benign tissue as also being cancerous. Different users reacted differently to this, depending on their mental model of the AI's design objective.

To some participants, the AI's objective was to be as accurate as possible, independent of its end-user. These participants quickly lost trust in the AI when they observed that it fell short of being a gold standard: *"Their grading of 4 is wrong, I would say forget about this, give me a clean slide and I'll make my own decision...I cannot trust this grading at all."* (P16). To others, however, the AI's role was to merely draw their attention to suspicious regions, given that the pathologist will be the one to make sense of those regions anyway: *"It just gives you a big picture of this is the area it thinks is suspicious. You can just look at it and it doesn't have to be very accurate."* (P14) Some compared the AI's predictions to the help they typically get from medical residents, who make rough mark-ups of questionable regions for them to review. Rather than expecting pixel-perfect predictions, these pathologists interpreted the AI's objective as that of drawing attention to worrisome regions, which a human will ultimately interpret. These user expectations surrounding the imperfect nature of annotations are consistent with prior research on how medical experts communicate decision thought processes to one another through lightweight, informal mark-ups [30]. As human–AI

decision-making becomes more prevalent, it may become even more crucial to make explicit the extent to which an algorithm's objective function accounts for the presence of a human collaborator.

6 DISCUSSION

In this study, we found that, beyond the need to explain specific model decisions, clinicians had broader desires to form an initial impression of a model's general tendencies, such as its specific limitations and pitfalls, its medical point-of-view and idiosyncrasies, and its overall design goals. These global properties of the system could help inform their interpretation of local events and predictions during critical decision-making. While this work reports on results from a study of pathologists, our findings are likely applicable to other contexts in which experts partner with AI Assistants, especially when high-stakes decisions are being made. In light of our findings, we now discuss the broader implications of this research, and how onboarding materials and the onboarding process could be designed to better support these information needs.

6.1 Domain-Specific, Human-Relatable Test Cases

Participants implicitly and explicitly understood that no tool (or person) is perfect. In part, this understanding is derived from their training and on-the-job experiences, both of which have supplied them with numerous examples of cases that can be challenging (e.g., benign mimickers of cancer). This knowledge of where they themselves can fail led to a desire to understand how a new tool will fare in these same situations. The relative ease with which participants were able to enumerate test cases suggests that it may be fairly straightforward to define a library of *domain-specific test cases* for an AI Assistant. These domain-specific test cases represent situations that experts consider to be challenging and potentially of critical importance, similar to the types of cases that senior mentors might select as training for a junior colleague or trainee [28, 68].

In part, these test cases represent a desire to understand the AI Assistant in human-relatable terms. Participants wished to understand how well the Assistant performs relative to them, or a panel of GU pathologists. Such test cases could be used during an interactive calibration phase of onboarding, during which a user diagnoses a small set of cases and observes the AI's performance, as well as that of GU pathologists. This form of interactive testing could offer additional color to raw performance numbers, while demonstrating the Assistant's capabilities in ways that users can immediately relate to.

A larger library of domain-specific test cases could also be leveraged to provide an overview of a model's strengths and weaknesses during onboarding, by communicating model performance on each category of cases prior to use. Performance on these categories may also be re-surfaced on-demand by users, for reference during diagnosis. For example, when the user's diagnosis differs from that of the Assistant, and if there are clear patterns in the Assistant's accuracy for a given situation, the user could take this knowledge into consideration when factoring the Assistant's prediction into their decision making process. As such, communicating performance on these test cases could also support competence articulation: similar to the human practice of articulating one's competence to collaborators [45], the AI Assistant could articulate its scope and expertise to the end-user through the report of categorical strengths and weaknesses. Articulating competence during onboarding is in line with recent human–AI guidelines [5], which recommend making AI system capabilities clear to users during initial interaction.

Eliciting important, domain-specific test cases may also have benefits when developing the underlying model and designing the user-facing portion of the AI Assistant. In the case of modeling, these test cases highlight classes of inputs that should be adequately represented in the training data, and appropriately handled by the model. For the design of the AI Assistant, these test cases

represent a set of use cases that can inform design choices, while providing specific cases to use when conducting formative and summative evaluations of the Assistant.

Our findings also reinforce the idea that it can be useful to engage stakeholders and end-users in early AI engineering efforts (e.g., in the spirit of participatory design or co-design [36]). For example, identifying domain-specific test cases may suggest training data needs, as well as cases useful for onboarding and for validating model behavior. Thus, a potentially fruitful area of future research is to explore additional ways stakeholders and end-users can inform early AI engineering efforts.

6.2 AI Primer

Pathologists currently use tools that are grounded in physical and biological phenomena; they have relatively little exposure to the specialized field of machine learning. As a result, they do not necessarily know what questions they *should* be asking about an AI Assistant. Similarly, they lack a basic understanding of how these systems operate, which can lead to confusion when the Assistant behaves in ways that do not align with physical and biological realities (e.g., when it splits biological structures in half, assigning them two separate grades). Accordingly, it may be useful to provide participants with a basic *AI Primer* during onboarding, to introduce users to the basic concepts and process of machine learning.

As an example, an AI primer could introduce users to the basic process of machine learning (e.g., that it learns through pattern recognition and iterative tuning from processing many examples). For a specific Assistant, it could also explain the extent to which that particular model could have learned higher-order concepts from raw pixels (e.g., it may recognize the visual patterns of biological entities, but has not been explicitly trained on textbook biological knowledge and pathways). To caution users upfront about AI-specific behavior that may be surprising, pilot users could be provided something akin to a Turing Test, where their goal is to distinguish between AI behavior that is fathomable versus AI quirks that even an inexperienced clinician would not display. These quirks can then be addressed in the AI Primer.

6.3 Beyond Accuracy: Communicate the AI's Point-of-View and Design Goal

As we have indicated, we found that knowledge of the AI's accuracy was only one desired characterization of the AI's capabilities. Much in the way they might assess a colleague when considering their opinions, participants desired to evaluate the AI not only in terms of its competence, but also its subjective style (e.g., how conservative or liberal its decisions are, or its specific diagnostic patterns). Moreover, users' trust in an AI counterpart and expected mode of collaboration can depend on their mental model of its value criteria (e.g., whether it places more emphasis on sensitivity at the cost of specificity), and the design decisions that went into its creation (e.g., whether the goal of complementing strengths and weaknesses of a human counterpart were baked into its objective function). Hence, in high-stakes decision-making, where there is often grey area between decisions, providing users with the AI's world view may be just as crucial as providing its accuracy. For example, given a curated set of grey area cases labeled by a panel of GU pathologists, an interactive *human–AI calibration phase* could be useful. In this activity, both the pathologist and the AI could grade the cases, allowing them to quickly see for themselves how their subjective thresholds differ from the AI, as well as how they differ from a sample of experts. In other cases, global design decisions can be documented during development and fed into onboarding materials. Overall, these findings build on and extend research on value-sensitive algorithm design [82]: For example, an AI may be designed to place a higher value on sensitivity than specificity; on being somewhat conservative to avoid unnecessary treatment; or on minimizing certain classes of clinical errors at the expense of others. In addition to incorporating human values into algorithmic

design, algorithms should also make these value criteria and objectives transparent to end-users. Communicating the inherent values of AI to end-users will be important future work.

6.4 Onboarding to Modern-Day AI Assistants

On the one hand, our findings echo key information needs surfaced in a rich body of prior CAD research: as with first-generational CAD tools, there is a need to understand the scope and limitations of an AI system [32, 34], the potentially counter-intuitive errors it may make [31, 32], and its specific intended use [3].

Beyond this, our study also uncovered new challenges and needs associated with modern-day machine learning, and suggests concrete ways that onboarding could help address these needs. For example, some participants desired to know the quantity and diversity of the training data so that they could get a sense of the AI's capability and generalizability to their own settings (section 5.1.3). This finding relates to earlier work on the collaborative use of medical databases in which users desired to understand others' local data and contexts [37]. In our case, users desire this context to better understand the capabilities of a data-driven decision-making AI. These data-specific questions suggest that, as end-users become more familiar with the concept of ML learning from examples, data provenance will need to be considered when designing for onboarding.

The shift from detection-based systems toward more interpretive, diagnostic-based systems also raises new questions and considerations when designing for onboarding. In our study, users frequently interpreted AI assistants through the lens of their own collaborative practices with colleagues and medical training. For example, when shown diagnostic predictions, participants reflected on the subjectivity inherent in medical diagnosis, and thus felt it critical to know how ground truth was determined (section 5.3.2). In many ways, and perhaps more than ever before, pathologists were making sense of this new technology in terms of their own clinical practices, norms, and social dynamics, considering the "clinical perspective" of the AI much as they would a colleague. As the technical capabilities of AI continue to evolve and reach human performance levels, considering AI through the lens of existing professional and interpersonal practices may become increasingly pertinent. Simultaneously, these observations also raise theoretical questions around how human-AI onboarding ought to be studied: since people seem to think of these assistants as both tools and a partial stand-in for a human collaborator, it could be useful to adopt several theoretical lenses (e.g., tool-centric and human-centric) when studying these systems.

Finally, pathologists expressed an awareness that AI systems can be tuned to optimize for a variety of different competing priorities and use cases; for example, those who had used pap smear screening tools had developed an awareness that these tools may favor sensitivity at the expense of specificity. Likewise, they wondered if such design trade-offs were made with the diagnostic AI used in our study. Relative to detection-based CAD tools, diagnostic CAD may take on a much wider range of human-AI partnership modalities (e.g. pre-screen, pre-read, second opinion, etc.), making it perhaps even more critical to divulge the tool's design objective and intended use (section 5.4). In addition, user expectations of an AI's performance and metrics can be strongly anchored to their prior experience with detection-based models (section 5.1.1); as seen in our study, this can sometimes yield unrealistic expectations. Thus, relative to early CAD research, modern-day ML onboarding may need to take into consideration the evolving body of experience that clinicians have with CAD systems, and the assumptions they may have developed of AI through this process.

6.5 Onboarding Over Time

Our study focused on users' initial impressions of what information they felt necessary to effectively collaborate with an AI Assistant. However, one's relationship with a tool evolves over time. Likewise, the tool itself can change. In the case of an AI Assistant, its overall performance can change (e.g.,

the underlying model may be routinely re-trained, or clinical guidelines may change), even if its user interface remains the same. While this paper has considered onboarding for first-time use, an open question is what users' information needs are after installation of a new model.

One goal of onboarding materials is to educate users about the most effective way to use the tool (e.g., to indicate in which situations it excels, and in which situations it may be less reliable). The extent to which users can internalize and reliably apply this information is unclear, and is likely dependent on how detailed and nuanced the recommendations of use are. As such, it may be worthwhile to routinely refresh users' knowledge on the most effective way to use the Assistant, to increase the chance that they attain the most desirable outcome over time. In this light, we view onboarding as a process that is critical for first-time use, but also one that unfolds over time.

6.6 Future Work

The findings from this work provide a foundation for further research focused on the initial, introductory phases of using an AI Assistant for collaborative decision making. Most immediately, there are clear opportunities to design and test onboarding materials based on the findings of this research. For example, there are open questions of how AI Assistant onboarding materials can shape work practices (e.g., to help people develop more effective strategies faster); how they can instill more accurate and actionable mental models; and how these materials impact assessments and attitudes toward the AI Assistant (e.g., what are the effects on user trust).

7 CONCLUSION

In this study, we have described the information needs of medical experts during their introduction to an AI Assistant. While substantial work has been performed on the topic of explaining model predictions, this study suggests that it could also be useful to provide transparency into the higher-level design objectives of the model itself, as well as its global behavior and tendencies. Future work should investigate how to efficiently design onboarding materials based on our results, to deliver this desired information in practice.

AI Assistants are likely to become more common fixtures of work in years to come. Our research findings contribute to the growing literature that examines how to make these systems useful, usable, and understandable partners in conducting work, specifically in the case of high stakes, expert decision making, an area that will likely continue to rise in importance in the near future.

8 ACKNOWLEDGMENTS

We thank Davis Foote, Kunal Nagpal, Craig Mermel, Adam Pearce, Pan-Pan Jiang, Rory Sayres, Isabelle Flament, Trissia Brown, Susan Huang, Jason Hipp, Abigail Huang, Ben Wedin, Andrei Kapishnikov, Jimbo Wilson, Matthew Symonds, Dan Russell, Fernanda Viegas, and Martin Wattenberg for their valuable help and thoughtful feedback on this research.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[2] Joanna Abraham and Madhu C Reddy. 2013. Re-coordinating activities: an investigation of articulation work in patient transfers. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 67–78.

[3] Eugenio Alberdi, AA Povyakalo, Lorenzo Strigini, Peter Ayton, Mark Hartswood, Rob Procter, and Roger Slack. 2005. Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation. *The British journal of radiology* 78, suppl_1 (2005), S31–S40.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for Human-AI Interaction. (2019).

[6] Hidetaka Arimura, Chiaki Tokunaga, Yasuo Yamashita, and Jumpei Kuwazuru. 2012. Magnetic resonance image analysis for brain CAD systems with machine learning. In *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*. IGI Global, 258–296.

[7] Jakob E Bardram and Thomas R Hansen. 2010. Context-based workplace awareness. *Computer Supported Cooperative Work (CSCW)* 19, 2 (2010), 105–138.

[8] Pernille Bjørn and Morten Hertzum. 2011. Artefactual multiplicity: A study of emergency-department whiteboards. *Computer Supported Cooperative Work (CSCW)* 20, 1-2 (2011), 93–121.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Virginia Braun and Victoria Clarke. 2014. What can "thematic analysis" offer health and wellbeing researchers? *International journal of qualitative studies on health and well-being* 9 (2014).

[11] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. 2019. Automated Gleason Grading of Prostate Biopsies using Deep Learning. *arXiv preprint arXiv:1907.07980* (2019).

[12] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[13] Ayşe G Büyüktür and Mark S Ackerman. 2017. Information Work in Bone Marrow Transplant: Reducing Misalignment of Perspectives. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 1740–1752.

[14] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 258–262.

[15] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, and Michael Terry. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 4.

[16] Peter Carruthers and Peter K Smith. 1996. *Theories of theories of mind*. Cambridge University Press.

[17] Sharolyn Converse, JA Cannon-Bowers, and E Salas. 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues* 221 (1993).

[18] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[19] Paul Dourish. 2003. The appropriation of interactive technologies: Some lessons from placeless documents. *Computer Supported Cooperative Work (CSCW)* 12, 4 (2003), 465–490.

[20] Jonathan I Epstein, Michael J Zelefsky, Daniel D Sjoberg, Joel B Nelson, Lars Egevad, Cristina Magi-Galluzzi, Andrew J Vickers, Anil V Parwani, Victor E Reuter, Samson W Fine, et al. 2016. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *European urology* 69, 3 (2016), 428–435.

[21] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal* 1341, 3 (2009), 1.

[22] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 432.

[23] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[24] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[25] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[26] Maryellen L Giger. 2018. Machine learning in medical imaging. *Journal of the American College of Radiology* 15, 3 (2018), 512–520.

[27] Robert Greenes (Ed.). 2014. *Clinical Decision Support*. Academic Press.

[28] M Hartswood, L Blot, P Taylor, S Anderson, R Procter, L Wilkinson, and L Smart. 2009. Reading the lesson: eliciting requirements for a mammography training application. In *Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment*, Vol. 7263. International Society for Optics and Photonics, 72631D.

[29] Mark Hartswood, Marina Jirotka, Rob Procter, Roger Slack, Alex Voss, and Sharon Lloyd. 2005. Working IT out in e-Science: Experiences of requirements capture in a HealthGrid project. *Studies in health technology and informatics* 112 (2005), 198–209.

[30] Mark Hartswood, Rob Procter, Mark Rouncefield, and Roger Slack. 2002. Performance management in breast screening: A case study of professional vision. *Cognition, Technology & Work* 4, 2 (2002), 91–100.

[31] Mark Hartswood, Rob Procter, Mark Rouncefield, Roger Slack, James Soutter, and Alex Voss. 2003. 'Repairing' the Machine: A Case Study of the Evaluation of Computer-Aided Detection Tools in Breast Screening. In *ECSCW 2003*. Springer, 375–394.

[32] Mark Hartswood, Rob Procter, and L Williams. 1998. Prompting in mammography: Computer-aided Detection or Computer-aided Diagnosis. *Proceedings of Medical Image Understanding and Analysis, MIUA* 98 (1998), 6–7.

[33] Mark Hartswood, Rob Procter, Linda Williams, Robin Prescott, and Pat Dixon. 1996. Subjective responses to prompting in screening mammography. *MIUA-96* (1996).

[34] Mark Hartswood, R Procter, L Williams, R Prescott, and P Dixon. 1997. Drawing the line between perception and interpretation in computer-aided mammography. In *Proceedings of the First International Conference on Allocation of Functions*. Citeseer, 275–291.

[35] 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), 56.

[36] Ken Hinckley (Ed.). 2018. *ACM Trans. Comput.-Hum. Interact.* 25, 1 (2018).

[37] Marina Jirotka, Rob Procter, Mark Hartswood, Roger Slack, Andrew Simpson, Catelijne Coopmans, Chris Hinds, and Alex Voss. 2005. Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work (CSCW)* 14, 4 (2005), 369–398.

[38] Bridget Kane and Saturnino Luz. 2009. Achieving diagnosis by consensus. *Computer Supported Cooperative Work (CSCW)* 18, 4 (2009), 357–392.

[39] Zvi Kaufman, Baruch Shpitz, Myra Shapiro, Ronny Rona, Sylvia Lew, and Alex Dinbar. 1994. Triple approach in the diagnosis of dominant breast masses: Combined physical examination, mammography, and fine-needle aspiration. *Journal of surgical oncology* 56, 4 (1994), 254–257.

[40] Brian Keeffe, Usha Subramanian, William M Tierney, Edmunds Udris, Jim Willems, Mary McDonell, and Stephan D Fihn. 2005. Provider response to computer-based care suggestions for chronic heart failure. *Medical care* (2005), 461–465.

[41] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR medical informatics* 6, 2 (2018).

[42] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279* (2017).

[43] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1885–1894.

[44] Ajay Kohli and Saurabh Jha. 2018. Why CAD failed in mammography. *Journal of the American College of Radiology* 15, 3 (2018), 535–537.

[45] Simon B Larsen and Jakob E Bardram. 2008. Competence articulation: alignment of competences and responsibilities in synchronous telemedical collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 553–562.

[46] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3365–3376.

[47] Jennifer Marie Logg. 2016. *When do people rely on algorithms?* Ph.D. Dissertation. UC Berkeley.

[48] Thomas M. Maddox, John S. Rumsfeld, and Philip R. O. Payne. 2019. Questions for Artificial Intelligence in Health Care. *JAMA* 321, 1 (01 2019), 31–32. https://doi.org/10.1001/jama.2018.18932 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2718456/jama_maddox_2018_vp_180150.pdf

[49] Gregory Maniatopoulos, Rob Procter, Sue Llewellyn, Gill Harvey, and Alan Boyd. 2015. Moving beyond local practice: reconfiguring the adoption of a breast cancer diagnostic technology. *Social Science & Medicine* 131 (2015), 98–106.

[50] Helena M Mentis. 2017. Collocated Use of Imaging Systems in Coordinated Surgical Practice. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 78.

[51] 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*. ACM, 220–229.

[52] Jesper Molin, Paweł W Woźniak, Claes Lundström, Darren Treanor, and Morten Fjeld. 2016. Understanding design for automated image analysis in digital pathology. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM, 58.

[53] Katherine T Morris, Rodney F Pommier, Arden Morris, Waldemar A Schmidt, Gregory Beagle, Priscilla W Alexander, SuEllen Toth-Fejel, Josh Schmidt, and John T Vetto. 2001. Usefulness of the triple test score for palpable breast masses. *Archives of Surgery* 136, 9 (2001), 1008–1013.

[54] Clara Mosquera-Lopez, Sos Agaian, Alejandro Velez-Hoyos, and Ian Thompson. 2015. Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. *IEEE reviews in biomedical engineering* 8 (2015), 98–113.

[55] Mark A Musen, Blackford Middleton, and Robert A Greenes. 2014. Clinical decision-support systems. In *Biomedical informatics*. Springer, 643–674.

[56] 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), 48.

[57] Tan Huu Nguyen, Shamira Sridharan, Virgilia Macias, Andre Kajdacsy-Balla, Jonathan Melamed, Minh N Do, and Gabriel Popescu. 2017. Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *Journal of biomedical optics* 22, 3 (2017), 036015.

[58] Sharoda Aurushi Paul. 2009. Understanding together: sensemaking in collaborative information seeking. (2009).

[59] Josefin Persson, Ulrica Wilderäng, Thomas Jiborn, Peter N Wiklund, Jan-Erik Damber, Jonas Hugosson, Gunnar Steineck, Eva Haglind, and Anders Bjartell. 2014. Interobserver variability in the pathological assessment of radical prostatectomy specimens: findings of the Laparoscopic Prostatectomy Robot Open (LAPPRO) study. *Scandinavian journal of urology* 48, 2 (2014), 160–167.

[60] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[61] Madhu Reddy and Paul Dourish. 2002. A finger on the pulse: temporal rhythms and information seeking in medical work. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 344–353.

[62] Madhu C Reddy, Paul Dourish, and Wanda Pratt. 2006. Temporality in medical work: Time also matters. *Computer Supported Cooperative Work (CSCW)* 15, 1 (2006), 29–53.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[64] Steven Schalekamp, Bram van Ginneken, Cornelia M Schaefer-Prokop, and Nico Karssemeijer. 2014. Influence of study design in receiver operating characteristics studies: sequential versus independent reading. *Journal of Medical Imaging* 1, 1 (2014), 015501.

[65] Peter G Scupelli, Yan Xiao, Susan R Fussell, Sara Kiesler, and Mark D Gross. 2010. Supporting coordination in surgical suites: Physical aspects of common information spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1777–1786.

[66] Korsuk Sirinukunwattana, Shan e Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* 35, 5 (2016), 1196–1206.

[67] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[68] James Soutter, Joao Campos, Mark Hartswood, Marina Jirotka, Rob Procter, Roger Slack, and Paul Taylor. 2003. Grid-based mammography training. *Hospital Radiologist* 5, 6 (2003).

[69] Peter Ström, Kimmo Kartasalo, Henrik Olsson, Leslie Solorzano, Brett Delahunt, Daniel M Berney, David G Bostwick, Andrew J Evans, David J Grignon, Peter A Humphrey, et al. 2019. Pathologist-Level Grading of Prostate Biopsies with Artificial Intelligence. *arXiv preprint arXiv:1907.01368* (2019).

[70] Jie Tan, Matthew Ung, Chao Cheng, and Casey S Greene. 2014. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In *Pacific Symposium on Biocomputing Co-Chairs*. World Scientific, 132–143.

[71] Hilda Tellioğlu and Ina Wagner. 2001. Work practices surrounding PACS: the politics of space in hospitals. *Computer Supported Cooperative Work (CSCW)* 10, 2 (2001), 163–188.

[72] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. 2018. Machine learning in medicine: Addressing ethical challenges. *PLoS medicine* 15, 11 (2018), e1002689.

[73] Sergio G Veloso, Mario F Lima, Paulo G Salles, Cynthia K Berenstein, Joao D Scalon, and Eduardo A Bambirra. 2007. Interobserver agreement of Gleason score and modified Gleason score in needle biopsy and in surgical specimen of prostate cancer. *International braz j urol* 33, 5 (2007), 639–651.

[74] Abraham Verghese, Nigam H. Shah, and Robert A. Harrington. 2018. What This Computer Needs Is a Physician: Humanism and Artificial Intelligenc. *JAMA* 319, 1 (01 2018), 19–20. https://doi.org/10.1001/jama.2017.19198 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2666717/jama_verghese_2017_vp_170180.pdf

[75] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. (2019).

[76] Robert L Wears and Marc Berg. 2005. Computer technology and clinical work: still waiting for Godot. *Jama* 293, 10 (2005), 1261–1263.

[77] Daniel M Wegner, Toni Giuliano, and Paula T Hertel. 1985. Cognitive interdependence in close relationships. In *Compatible and incompatible relationships*. Springer, 253–276.

[78] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. (2019).

[79] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. (2019).

[80] A Zamora, F Fernández de Bobadilla, C Carrion, G Vázquez, G Paluzie, R Elosua, M Vilaseca, A Martín-Urda, A Rivera, N Plana, et al. 2013. Pilot study to validate a computer-based clinical decision support system for dyslipidemia treatment (HTE-DLP). *Atherosclerosis* 231, 2 (2013), 401–404.

[81] Zhan Zhang and Aleksandra Sarcevic. 2018. Coordination Mechanisms for Self-Organized Work in an Emergency Communication Center. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 199.

[82] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 194.

Received April 2019; revised June 2019; accepted August 2019