Ambiguity-aware AI Assistants for Medical Data Analysis

Mike Schaekermann[1], **Graeme Beaton**[1], **Elaheh Sanoubari**[1],
Andrew Lim[2], **Kate Larson**[1], **Edith Law**[1]
[1]University of Waterloo, [2]University of Toronto, Canada
{mschaeke,graeme.beaton,elaheh.sanoubari}@uwaterloo.ca,
andrew.lim@utoronto.ca, {kate.larson,edith.law}@uwaterloo.ca

ABSTRACT
Artificial intelligence (AI) assistants for clinical decision making show increasing promise in medicine. However, medical assessments can be contentious, leading to expert disagreement. This raises the question of how AI assistants should be designed to handle the classification of ambiguous cases. Our study compared two AI assistants that provide classification labels for medical time series data along with quantitative uncertainty estimates: conventional vs. *ambiguity-aware*. We simulated our ambiguity-aware AI based on real-world expert discussions to highlight cases likely to lead to expert disagreement, and to present arguments for conflicting classification choices. Our results demonstrate that ambiguity-aware AI can alter expert workflows by significantly increasing the proportion of contentious cases reviewed. We also found that the relevance of AI-provided arguments (selected from guidelines either randomly or by experts) affected experts' accuracy at revising AI-suggested labels. Our work contributes a novel perspective on the design of AI for contentious clinical assessments.

Author Keywords
Ambiguity; Artificial Intelligence; Medical Data Analysis.

CCS Concepts
•Human-centered computing → Human computer interaction (HCI);

INTRODUCTION
AI systems show increasing promise for numerous clinical applications. Recent advances in deep learning have spawned AI systems with expert-level performance in several domains of medical data classification (e.g., [43, 44, 57]). However, contentious patient cases leading to expert disagreement are prevalent in medicine [32]. Given the gravity of correct clinical assessments, an important question in the design of AI for medical data analysis is how the system should communicate uncertainty about the classification of ambiguous cases.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04.
http://dx.doi.org/10.1145/3313831.3376506

State-of-the-art AI systems are capable of providing quantitative uncertainty estimates (e.g., 70% confident that a patient case is abnormal). These estimates are typically derived from posterior probability distributions over the possible classification labels. However, prior work has shown that these estimates do not always reliably predict expert disagreement [42]. Furthermore, numeric representations of uncertainty alone may not be sufficient for human experts to make sense of the underlying reasons behind the AI's uncertainty.

Prior work in explainable AI (XAI) has established the importance of providing reasons for AI-suggested labels to foster model transparency and user trust [1, 41, 62]. Building on this body of work, we argue that explanations for label ambiguity can be leveraged by AI assistants to support medical reasoning. We detail a within-subject study with twelve expert participants who interacted with both a conventional and an *ambiguity-aware* AI assistant, reviewing a total of 4,514 AI-suggested labels, out of which 22% were contentious. Both assistants used quantitative representations to communicate uncertainty, but our ambiguity-aware AI also highlighted contentious cases and explained why they were ambiguous by providing human-interpretable arguments for the conflicting labels. While this feature was simulated using cases and arguments selected from real-world expert discussions, participants were unaware of its simulated nature. Our findings suggest that explaining ambiguity can benefit AI-assisted medical reasoning. Our main contributions are:

1. We present a novel approach for communicating ambiguity in AI-assisted medical reasoning, and provide evidence that ambiguity-aware AI can alter experts' workflows by effectively re-directing their attention and review activity to contentious cases.

2. We demonstrate that while explaining ambiguity can contribute to experts' labeling accuracy, its impact heavily depends on the relevance of the arguments provided (selected from guidelines either randomly or by experts). Specifically, if the arguments are not sufficiently relevant, experts' accuracy can suffer to the point below that of random guessing (i.e., less than 50% accurate).

3. We provide design considerations for communicating uncertainty in AI-assisted medical reasoning, laying a foundation for future implementations of AI systems better capable of conveying information about contentious cases.

In the following sections, we outline related work on the issue of ambiguity and expert disagreement in medicine, approaches for handling ambiguity in AI systems, clinical decision support technology, and the relationship between explainability and trust in AI systems. We then introduce the design of our AI assistants, followed by our research questions, hypotheses and methods. Finally, we detail our quantitative and qualitative findings, and conclude with a discussion of design considerations.

RELATED WORK

Ambiguity & Expert Disagreement in Medicine

Expert disagreement in medical data analysis has been deemed a "full-fledged clinical problem" [42]. There are various reasons for inter-rater disagreement in data classification tasks. Experts may disagree about a classification decision due to ambiguous problem definitions [2, 11, 23, 34], ambiguity inherent in the data itself [49], or the existence of more than one correct answer [15, 45].

Prior work in medical decision making describes that medical experts are susceptible to biases in their reasoning; for instance, "confirmation bias" can lead a medical expert to look only for evidence that is in line with their pre-existing hypothesis [6]. As sub-optimal decision-making in medicine can have major consequences, it is crucial to combat any reasoning biases medical experts may have. Our simulated ambiguity-aware AI aims to mitigate this bias by putting forth arguments for conflicting medical assessments, encouraging perspective-taking for alternate lines of reasoning.

Related literature suggests that communicating uncertainty can impact cognition and trust, and potentially influence experts' decision-making behaviours [60]. That said, there is a body of work showing that people have a general aversion towards ambiguity [30, 59]. For example, a study by Redelmeier and Shafir suggested that the uncertainty between two medical assessments led some doctors to avoid making a decision altogether [45]. Work done in psychology acknowledges ambiguity-tolerance as a personality variable [7, 26]. Medical education research advocates that given the inevitable nature of uncertainty in contemporary medicine, medical experts must acquire a certain level of tolerance to it [33].

Handling Ambiguity in AI Systems

Prior systems have generally taken one of three approaches to the problem of ambiguity in AI-based data classification:

Eliminating Ambiguity. Traditional machine learning classification methods eliminate class diversity using automatic procedures like majority vote [28], or expectation maximization [38]. These systems tend to view ambiguity as a proxy for noise to be reduced or eliminated in the data. [10, 63].

Aggregating Multiple Outputs. Other systems retain disagreement labels for the purpose of training multiple models (e.g., one for each human labeler [22]); these systems typically produce multiple AI predictions which are aggregated into a single label before being presented to the end user.

Label Distribution Learning. A more ambiguity-centric approach to data classification is label distribution learning (LDL) [20], where machines are trained to predict not just one label for a given case, but a distribution of possible classification labels [13, 40]. Standard LDL models will assign uncertainty estimates to their classification outputs, providing degrees of plausibility for each possible label.

The question of how systems should *communicate* or visually represent uncertainty to end users has received ample attention in the human-computer interaction (HCI) community [56]. Approaches include visualizing uncertainty as extrinsic annotation (e.g., confidence intervals), abstract, continuous outcomes (e.g, probability density plots), or hypothetical, discrete outcomes (e.g., natural frequencies or icon arrays) [29]. Kay et. al [29] suggest that communicating uncertainty through discrete outcomes can improve decision making on the part of end users.

Prior work has found that collecting explanations around ambiguous cases during data labeling workflows can be leveraged towards more fine-grained and flexible post-hoc data classification [12]. In the context of medical data analysis, Schaekermann et al. [50] showed that discussion metadata produced by medical specialists can be re-used for the training of medical generalists to better calibrate their grading approach for difficult cases. Galdran et al. [19] developed a system for vessel classification from retinal images, with the ability to classify uncertain cases and provide direct uncertainty estimates for its labels while achieving state-of-the-art classification performance.

We take inspiration from Galdran's work by simulating an ambiguity-aware AI assistant for medical data analysis, an LDL system that provides human-interpretable rationales for all plausible classification labels. While our AI assistant is simulated in the sense that it does not predict, but merely displays human-annotated data, our work contributes novel insights about how such an ambiguity-aware system affects expert perception and behaviour.

Clinical Decision Support

Clinical decision support (CDS) is broadly defined as the provision of intelligent assistance to clinicians, medical staff, and patients [37]. CDS can include low-level functions like computerized alerts and reminders for providers and patients, or high-level functions like patient diagnosis [14]. Norman et al. [36] describe a dual process of diagnostic reasoning, where physicians engage in (1) a non-analytic or unconscious process of hypothesis generation, and/or (2) a conscious, analytic process of hypothesis testing. The latter is an extensive computational process, and has motivated efforts to develop AI-based CDS systems for diagnostic support.

In such support systems, a physician cross-checks the algorithmic output against their internal knowledge, but takes responsibility for the final diagnostic decision. Our work takes a similar approach of augmenting, instead of automating, the job of physicians [21].

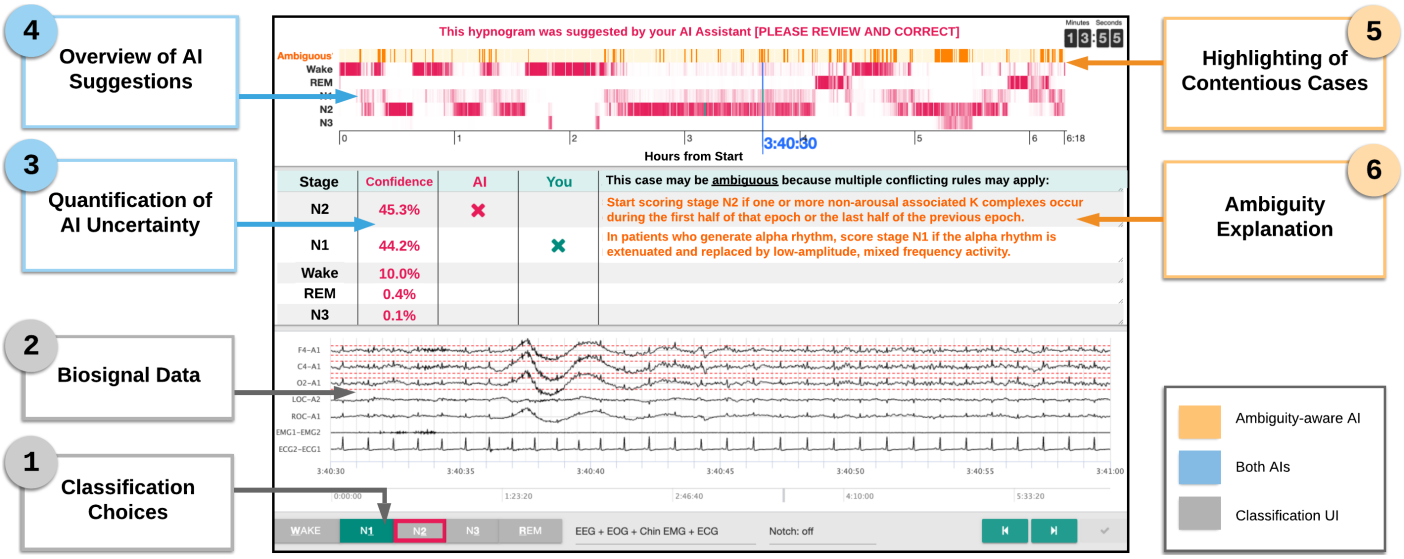

Figure 1. Interface for conventional and ambiguity-aware AI assistants in medical data analysis.

Barriers to the Adoption of AI-Based CDS Systems

Explainability. ML-based AI systems are typically opaque with respect to their internal functions [35]. In fields where AI is tasked with important decisions, it is imperative that automated decision making be interpretable, especially if the AI is known to be imperfect [8, 31]. The field of XAI [61] emerged as a response to this problem of transparency beginning in the 1970s and 1980s with the deployment of expert systems with explanation capabilities—most notably for medical decisions [1]. Explanations have been found to promote transparency in machine learning algorithms and make users more aware of how a system works [41]. Recent approaches in XAI, e.g., Ehsan et al. [18], demonstrate that AI systems can learn to generate human-like natural language explanations for their decisions. Mittelstaedt et al. [35] argue that there is a mechanistic link between explanation and justification in human discourse, and that machine explanations should emulate human explanations. Our simulated AI assistant instantiates these design principles, by providing human-interpretable rationales for its outputs.

Trust. A lack of trust is arguably the most significant barrier to adoption of AI-based systems. A CDS system can bias a physician to choose the wrong course of action against their own clinical judgement [17]. Human experts may also fail to trust a reliable system. It is crucial that an appropriate level of trust in automation be established to balance over-reliance and under-reliance. Cai et al. [8] demonstrated a link between explainability and trust by showing that pathologists trusted a CDS tool for cancer diagnosis more if they could tweak its internal representation of image similarity using domain-specific concepts (e.g., number of fused glands).

Addressing the problem of trust becomes more complicated in the context of uncertainty. Psychological uncertainty is an aversive state [60], and thus information must be communicated effectively to hedge against its negative effects. There appears to be a volatile relationship between uncertainty and trust. On the one hand, trust can be undermined by failing to communicate uncertainty; on the other hand, admitting uncertainty can also hinder trust [60]. Thus, it is crucial that machines strike the right balance between communicating and withholding uncertainty information.

Studies on how communicating uncertainty affects trust are limited and have produced mixed results. While there is some evidence that trust can be fostered through explained uncertainty [29], more research is needed. In a recent review of the matter, van der Bles [60] acknowledged that uncertainty does not always produce negative emotional effects. Indeed, in the healthcare domain, Schneider et al. [56] developed a system for communicating uncertainty in fertility prognosis that increased users' understanding of uncertainty without causing them to have a negative view of the system.

In this work, we study how the workflows and perception of medical experts is affected by an AI assistant capable of identifying and explaining ambiguous cases.

AMBIGUITY-AWARE AI ASSISTANCE

In this study, we explore how human-AI collaboration is affected by an AI system's ability to not only flag *if* specific edge cases are on the classification boundary between two or more categories, but also explain *why* a given case may be ambiguous. Specifically, we compare a simulated AI system that provides experts with arguments for conflicting classification choices for a contentious case to a conventional AI assistant that only provides numeric uncertainty estimates. Our ambiguity-aware AI system uses a Wizard of Oz approach. That is, justifications for conflicting classification labels were hand-authored by human experts using a round-based discussion procedure reported in prior work [49]. To compare the ambiguity-aware AI assistant to a conventional AI assistant, we led participants to believe that the justifications presented to them were generated by an AI while, in fact, they were manually selected by human experts.

Figure 1 illustrates how the two AI assistants—conventional AI vs ambiguity-aware AI—were integrated into an existing expert interface for classification of medical time series data. Both AIs suggested classification labels based on a state-of-the-art deep learning algorithm for sleep stage classification [57], which has an average accuracy of 87% (when judged against consensus labels from an expert panel). Both AI assistants provided a sequence overview of all suggested labels (hypnogram), in which each label corresponded to a 30-second segment in the timeline of a multi-hour patient recording. Experts could open a case by selecting the corresponding time window in the overview, or by navigating through the recording chronologically.

The key difference between the two AI assistants was in how they communicated uncertainty to expert end users. Typical output from machine learning algorithms includes not only the predicted classification label, but also a likelihood distribution over all possible classification choices. Both of our AI assistants were designed to communicate this type of *quantitative* uncertainty estimate in two ways (Figure 1, blue labels 3 and 4): (1) in the timeline overview, quantitative uncertainty was visualized by mapping the confidence level (in percentage) for each possible classification label to a *transparency* value used to display the label option in the timeline—low confidence classification labels were more transparent, and high confidence classification labels were more opaque; (2) in the case detail view, quantitative uncertainty was displayed in a tabular format, listing all possible classification choices ordered from most to least likely along with their *percentage* confidence levels.

While our conventional AI employed this baseline representation of uncertainty, our ambiguity-aware AI also communicated *qualitative* uncertainty based on arguments gathered from real-world expert discussions (Figure 1, orange labels 5 and 6). Specifically, the timeline overview was augmented with an additional layer **highlighting contentious cases** that were likely to spur expert disagreement. Note that these suggestions did not dictate the order in which cases were presented to experts for review: experts were still free to decide how to navigate the recording timeline and what cases to review in which order. In addition, the case detail view for contentious cases was extended with an **ambiguity explanation**, listing human-interpretable arguments for conflicting classification choices. These arguments corresponded to discrete scoring rules from the official guidelines for sleep stage classification, and were based on data from real-world expert discussions as described above.

RESEARCH QUESTIONS AND HYPOTHESES

Our work addresses two primary research questions about the impact of ambiguity-aware AI on the behaviour (Q1) and perception (Q2) of medical experts.

Q1: How does ambiguity-aware AI affect medical assessments?

Expert time is a limited and expensive resource in clinical settings and should therefore be allocated efficiently. We take the stance that while medical experts should make their clinical assessments with care, AI assistants can help prioritize which cases require their attention the most. Our ambiguity-aware AI is designed to redirect experts' attention towards cases likely to be contentious, and to provide arguments explaining the underlying classification ambiguity.

Our projection is that ambiguity explanations can inform clinical judgement and thus increase experts' classification accuracy without reducing the number of cases reviewed. Specifically, we envision that the relevance of ambiguity explanations is crucial for successfully informing expert judgement. We hypothesize that:

[H1a] The proportion of contentious cases reviewed by experts will be higher with an ambiguity-aware AI.

[H1b] Expert efficiency in terms of the overall number of cases reviewed will not suffer with an ambiguity-aware AI.

[H1c] Expert accuracy in terms of the overall portion of cases reviewed and labeled correctly will be higher with an ambiguity-aware AI.

[H1d] The accuracy of classification labels experts assigned to contentious cases will depend on the relevance of the provided ambiguity explanations.

Q2: How is ambiguity-aware AI perceived by medical experts?

HCI research has established that poor user perception can be a barrier to adoption of technology regardless of performance. It is therefore important to investigate expert perception, beyond the primary outcome of reliability in AI-assisted clinical assessments. We hypothesize that:

[H2a] Experts will have a preference for an ambiguity-aware AI.

[H2b] Experts will consider an ambiguity-aware AI more trustworthy.

[H2c] Highlighting and explaining contentious cases will not increase experts' cognitive load.

[H2d] Experts with higher ambiguity tolerance (as a personality trait) will have a stronger preference for the ambiguity-aware AI.

METHODS

Here we describe the details of our controlled experiment including the task, data set, study procedure, and statistical analysis. In our study, we simulate a scenario in which a medical AI assistant first analyzes a patient case to suggest classification labels of a certain kind. A trained medical expert then reviews and corrects as many AI-suggested classifications as possible within a given time window. This setting represents a future scenario where (imperfect) AI systems are deployed in time-sensitive clinical workflows while requiring oversight from human experts.

Task

We conducted our study in the field of biomedical time-series classification, an expert domain with typically high rates of

inter-rater disagreement. In particular, we compared our conventional and ambiguity-aware AIs in the context of assisting trained medical professionals in the task of sleep stage classification, i.e., analyzing a patient's sleep pattern based on medical time series data (polysomnograms) recorded in a sleep laboratory. A typical polysomnogram covers a whole night of sleep (i.e., six to eight hours).

Sleep technologists are responsible for physically recording sleep electroencephalograms (EEGs) by directly interacting with patients. They also annotate these recordings for physicians who then interpret and convey diagnoses to patients. Overnight sleep EEGs are widely used in the diagnosis of neurodegenerative and sleep disorders. The classification task used in our work is a key step in this process.

The task of sleep stage classification involves mapping fixed-length (30-second) segments of a polysomnogram to one of five sleep stages: Wake, Rapid Eye Movement (REM) sleep or one of three non-REM sleep stages (NREM 1, NREM 2, NREM 3). Figure 1 shows the expert classification interface used in our study. The resulting sequence of sleep stages, called a hypnogram, serves as evidence for the diagnosis of various neurological diseases and sleep-related disorders. Sleep technologists apply rules from official medical guidelines to classify time series segments into sleep stages based on visually inspecting the waveform patterns.

Sleep stage classification lends itself as a task for our study on AI assistance for contentious clinical assessments. Not only is it a time-consuming and tedious procedure; it also relies on lengthy and complex classification guidelines likely to spur expert disagreement. In fact, prior work has established that two sleep technologists have about a 17.4% chance of disagreeing on the correct classification of the same waveform segment [46].

Data

We selected two separate patient records (i.e, polysomnographic sleep studies) with similar characteristics (Table 1) to examine the two AI assistants under comparable conditions while avoiding learning effects on the side of experts.

	Patient A	**Patient B**
Pathology	Dementia	Dementia
Sex	Female	Male
Age Group	70-74 years	75-79 years
Recording Duration	6h 52 min 30 sec	6h 18 min 30 sec
# Cases Total	825	757
% Contentious Cases Total	18%	18%
% Contentious Cases out of all Correct AI Suggestions	12%	11%
% Contentious Cases out of all Incorrect AI Suggestions	48%	51%
AI Accuracy Overall	84%	83%
AI Accuracy on Contentious Cases	55%	51%

Table 1. Characteristics of patient records used by the AI assistants.

Note that patient records where selected such that the AI accuracy measured against just the contentious cases was close to 50% for both patients, meaning that correction by human experts was only required for about half of those cases. In addition to counter-balancing the order in which conventional and ambiguity-aware AI were presented to experts, the assignment of AI assistant to patient record was also fully counter-balanced. A separate third patient record was randomly selected for a practice phase preceding the main task.

Adjudication data. We source the data required to simulate our ambiguity-aware AI (i.e., which cases have high expert disagreement, and what are the arguments for different classification labels) from a previous study [49] on the use of group-based adjudication discussions in medical data analysis. This prior work introduced a round-based procedure to adjudicate clinical classification disagreements among groups of experts using a highly structured argument format. In particular, arguments were collected in the form of discrete classification rules taken from the official medical guidelines (e.g., *In patients who generate alpha rhythm, score stage N1 if the alpha rhythm is extenuated and replaced by low-amplitude, mixed frequency activity for more than half of the epoch*).

This data set was used to simulate output for our ambiguity-aware AI: cases that had caused expert disagreement and produced conflicting arguments in this data set were highlighted as contentious cases by our ambiguity-aware AI. Arguments put forward during the real-world adjudication process were presented for these cases to explain the ambiguity around conflicting classification labels. For Q1, we sought to examine the impact of argument relevance on clinical decision making for contentious cases (H1d). To this end, we added noise to ambiguity explanations by replacing a random subset (20%) of arguments with scoring rules randomly selected from the same medical guidelines. Otherwise, justifications were displayed as selected by experts during prior discussions, without further manipulation. Our randomization procedure was constrained to ensure that randomly selected arguments were never mentioned in the real-world expert discussion for a given case, and that all arguments presented were still pertinent to their classification choice: for example, an argument for REM sleep could only be replaced with another argument for REM sleep.

Finally, classification accuracy (of either AI or human experts) was measured against the consensus decision of our round-based adjudication procedure involving a panel of three independent experts for each classification decision.

Procedure

We recruited twelve sleep technologists as expert participants for our study. Our experts were recruited with the help of an allied sleep technologist from a local research clinic who posted our recruitment letter to a domain-specific Facebook group with about 4700 sleep technologists from different countries. Each expert was exposed to both AI assistants in a counter-balanced manner.

Consent procedure and pre-study survey. After providing informed consent for participation in the study, experts reported information about their demographics (age, gender, geographic location) and professional background (professional or academic training, number of years of professional experience). We employed the *Intolerance of Ambiguity* scale, a psychometric survey instrument developed by Budner [7], to learn about each expert's general level of tolerance for ambi-

guity in decision making. We included the phenomenological denial sub-scale consisting of four statements:

- *An expert who doesn't come up with a definite answer probably doesn't know too much.*
- *There is really no such things as a problem that can't be solved.*
- *People who insist upon a yes or no answer just don't know how complicated things really are.*
- *Many of our most important decisions are based on insufficient information.*

Experts rated their level of agreement for each of the four statements on a 7-point Likert scale.

Practice phase. Next, experts familiarized themselves for about 5 minutes with our waveform classification user interface and with the basic interface components common to both AI assistants.

Tasks. Experts performed the same main task twice, once with the ambiguity-aware AI assistant and once with the conventional variant, in a counter-balanced order. In each task, experts were asked to review the waveform of a particular patient record within a limited time window of 15 minutes. The patient record was fully pre-classified by the AI assistant and experts were asked to correct as many of the AI-suggested labels as possible within the given time limit. Experts could revise AI-suggested labels by selecting a different sleep stage label in the classification UI (Figure 1, gray labels 1 and 2). After each of the two tasks, experts filled out a brief feedback survey probing for their perception of each AI assistant. The survey included scales to measure perceived trust towards the AI assistant [27], cognitive load (NASA-TLX; [25]) during the task, perceived diagnostic utility and mental support provided by the AI assistant, and whether experts thought they would use the AI in practice.

Post-study survey. After completing the tasks, experts compared both AI assistants with respect to perceived reliability, trustworthiness, capability and provided an overall preference. Experts rated each of these four items on a 7-point Likert scale ranging from 1 (totally version A), 2 (much more version A than B), 3 (slightly more version A than B), 4 (neutral), etc. to 7 (totally version B). After completing the post-study survey, participants received a debrief statement informing them about the simulated nature of the ambiguity-aware AI in this study. Experts were compensated with CA$50 via online gift cards (or the equivalent amount in their preferred currency) for participation in the study, with an average study duration of one hour.

Analysis

For **Q1**, we investigated the impact of our ambiguity-aware AI on experts' behaviour in reviewing AI-suggested classification labels. We used dependent t-tests to compare both AI assistants with respect to the following outcome measures per expert: the proportion of contentious cases out of all reviewed cases (H1a), the number of cases reviewed given a fixed time window (H1b), and the accuracy rate of expert-provided labels (H1c). For our secondary analysis on the relevance of arguments for contentious cases (H1d), we used Pearson's chi-squared test of independence to compare experts' average accuracy at revising AI-suggested labels when presented with either expert-selected arguments only vs. cases with one or more randomly selected arguments.

For **Q2**, we compared experts' perception of both AI assistants. A possible trend in overall preference (H2a) for either of the AI assistants was examined using a one-sample Wilcoxon signed rank test. Self-reported scores for perceived trust (H2b) and cognitive workload (H2c) were compared between both AI assistants using Wilcoxon signed-rank tests. Finally, we used a Pearson's chi-squared test of independence to test whether experts' overall tendency of ambiguity tolerance (ambiguity-tolerant vs. intolerant) was associated with their overall preference for either AI assistant (preference for ambiguity-aware AI vs. conventional AI; H2d).

Finally, we used open coding to extract emerging themes from open-ended survey responses experts submitted after interacting with each AI. Experts were asked to reflect on how they decided which cases to review and why, what information they used to make these decisions, and how information about the AI's uncertainty affected their decision making. Recurring themes are reported below.

RESULTS

Expert Participants

Based on the pre-study questionnaire, our expert participants were located in the United States (6), Canada (4), the European Union (1) and one other unspecified location (1). Eleven of our expert participants reported having at least ten years of experience working as sleep technologists, and one participant reported having five to ten years of experience. Out of the twelve experts, five self-reported as female, six as male, and one participant did not specify their gender. The distribution over age groups was: 26-35 (1), 36-45 (7), 46-55 (1), 56+ (2), with one participant who did not specify their age group.

Q1: How does ambiguity-aware AI affect medical assessments?

We hypothesized that the ambiguity-aware AI assistant would alter experts' workflow and increase the number of contentious cases they review in the patient recording **(H1a)**. On average, the proportion of contentious cases out of all cases reviewed was significantly greater with the ambiguity-aware AI (M=.38, SE=.05) than with the conventional AI (M=.23, SE=.03), confirming our hypothesis (Figure 2). This difference was significant $t(11)=-2.82$, $p < .05$, indicating a large effect size r=.48.

We also hypothesized that using the ambiguity-aware AI would not negatively affect the number of cases reviewed by experts **(H1b)**. Our results show that there was no significant difference in the number of cases experts reviewed with the conventional AI (M=197.25, SE=47.60) compared with the ambiguity-aware AI (M=178.92, SE=57.02), $t(11)=.50$, p=.63. This result provides support for our hypothesis that experts' efficiency at reviewing AI-suggested labels was not

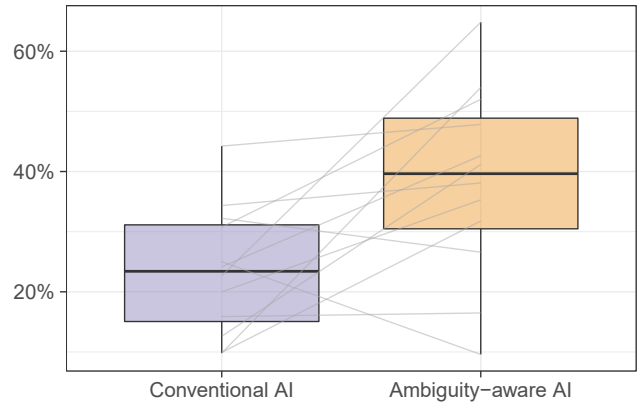

Figure 2. Proportion of contentious cases out of all cases reviewed. Ambiguity-aware AI guided experts' attention to contentious cases. Connecting lines correspond to individual experts.

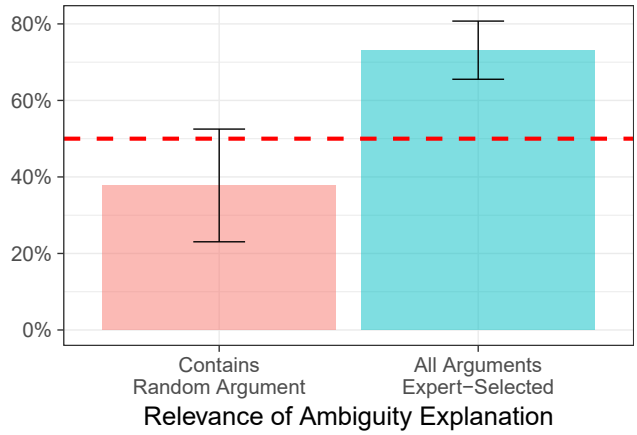

Figure 3. Experts' correction rate for cases with ambiguity explanation. The relevance of ambiguity explanations affects clinical assessments of contentious cases. Error bars present 95% confidence intervals.

negatively affected by being exposed to ambiguity explanations for contentious cases. Our projection that experts would achieve a higher overall labeling accuracy when assisted by the ambiguity-aware AI compared to the conventional one **(H1c)** could not be confirmed, t(11)=1.00, p=.34, r=.53.

Finally, we examined the potential impact of the relevance of ambiguity explanations for contentious cases on the likelihood that an expert would revise an AI-suggested label correctly **(H1d)**. We observed a significant association between the relevance of arguments (whether they contain randomly selected arguments or not) and experts' accuracy rate at revising AI suggestions χ^2=16.83, p < .001. In other words, the chance of a label getting revised correctly by an expert was significantly higher if the arguments provided were selected from guidelines via adjudication discussions (i.e., were relevant) than if they were selected from the guidelines randomly (Figure 3). The odds of a label getting revised correctly by an expert were 4.48 times higher (odds ratio) if the arguments provided were selected from guidelines by experts than if they were selected from the guidelines randomly.

Q2: How is ambiguity-aware AI perceived by medical experts?

For Q2, we explored experts' perception of both AI assistants. Results for our hypothesis that experts would have an overall preference for the ambiguity-aware AI **(H2a)** were mixed and were not statistically significant (p=.88). Except for two experts who did not have a preference for either AI, preferences were polarized. Out of the ten participants who expressed a preference, half preferred the ambiguity-aware AI assistant and the other half preferred the conventional AI (Figure 4).

While no significant differences could be detected between both AIs regarding perceived overall trust (p=.47), the ambiguity-aware variant was considered to have significantly greater integrity (p<.05), and we observed a trend that experts had higher confidence in the ambiguity-aware AI than in the conventional one (p=.09; Figure 5). These results provide partial support for our hypothesis **H2b**.

Furthermore, there were no detectably significant differences between the cognitive load scores of the two AI assistants on the NASA-TLX scale (p=.77), providing support for our hypothesis about their comparable mental demand **(H2c)**.

Finally, while experts varied in their level of ambiguity tolerance (M=17.25, SE=1.17), ranging from 10 to 26 on a scale from 4 to 28, no significant effect of ambiguity tolerance on expert perception could be detected (p=.62), leading us to reject hypothesis **H2d**.

Qualitative Insights

Our qualitative analysis of participant responses to open-ended survey questions yielded insights on how our ambiguity-aware AI assistant can affect experts' workflows and their mental model of AI assistants.

Altering expert workflows. Time constraints play an important role in real-world clinical workflows [58]. Case triaging—determining the priority for which cases receive an expert's attention first—is a common practice in medicine. Similarly, our ambiguity-aware AI assistant triages based on ambiguity by prioritizing contentious patient cases that need more attention from the expert.

Our qualitative findings suggest that some experts found the ambiguity-aware AI system to be more helpful in reducing cognitive load compared to the conventional assistant: *"Assistant B [ambiguity-aware] was more helpful in making me think as it listed the scoring rules that could apply to the epoch."*

Our analysis further highlights the effectiveness of the ambiguity-aware AI assistant in redirecting experts' attention to contentious cases. That is, six out of twelve experts in our study explicitly mentioned that their workflow differed between the two AI assistants, such that they prioritized checking contentious cases using the ambiguity-aware AI: *"I first*

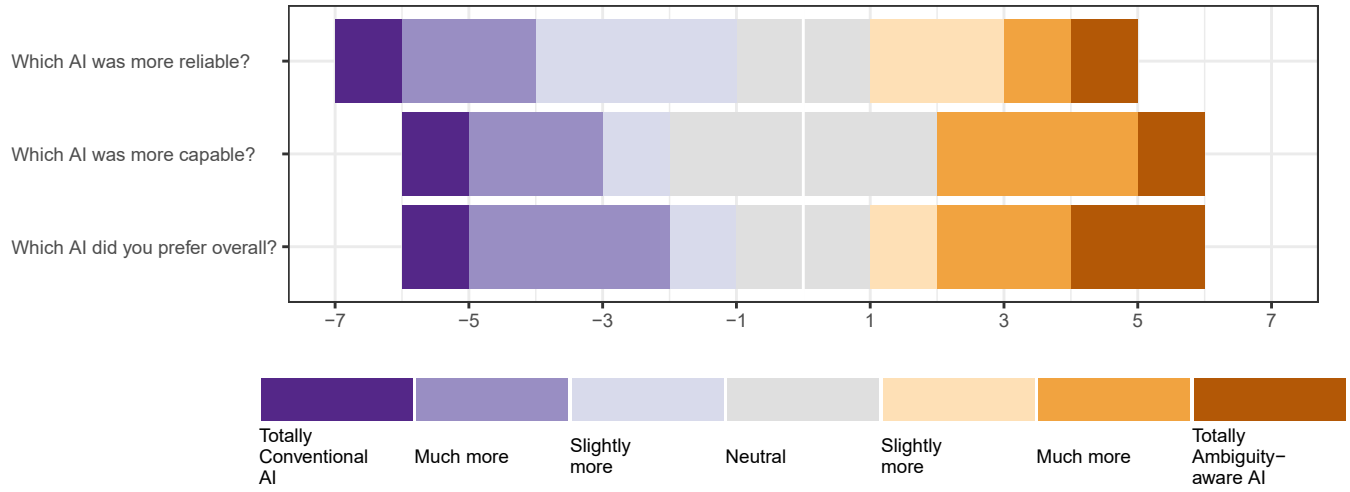

Figure 4. Experts' preferences between both AI assistants.

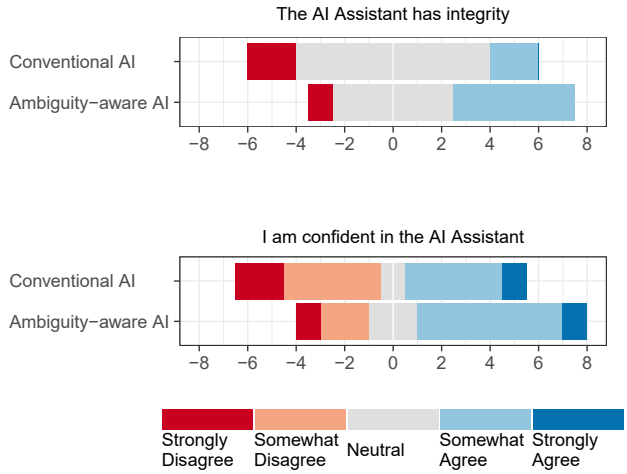

Figure 5. Expert ratings for perceived integrity and confidence from trust in automation scale.

chose the areas that the AI had marked as ambiguous and then tried to check sleep onset, REM onset, and stage 3 as time allowed."

One major criticism to the traditional approach of representing AI uncertainty with numeric confidence values is that it is not sufficient for experts to make sense of the underlying reasons behind the AI's uncertainty. Our qualitative evidence suggests that in choosing between numeric representations of AI uncertainty and human-interpretable ambiguity arguments experts found the latter to be more effective in guiding their attention: *"When I saw that the [conventional] AI had lower than an 80% confidence in the scored stage I tried to double check that epoch... I mostly used the areas marked as ambiguous [by the ambiguity-aware AI] as opposed to the percentage of certainty."*

In our study, we imposed time limits to understand how ambiguity-aware AI would help guide expert attention under the time constraints of real-world workflows. This temporal constraint was received differently by different expert participants. While some experts perceived the timers to be *"very frustrating"*, others found them useful: *"The time limit was great as my first instinct was to review the entire study and see if I was in agreement"*.

Mental models of AI assistants. Experts have preconceived mental models about the level of ambiguity in different cases. For instance, experts may draw from their prior experience of disagreements with other colleagues and have intuitions about what type of medical assessment is the most difficult to agree upon in their specific domain (e.g. certain classifications and stage transitions). It is therefore possible that these intuitions are projected onto the AI assistant to anticipate where the AI would likely make mistakes: *"I had to think where do we, as scoring techs, usually have the strongest disagreement and check those epochs."*

Beyond preconceptions, we also observed that experts developed comparative their mentals models about the two types of AI systems: *"AI 1 [ambiguity-aware] was rather impressive actually. Although in study 2, the persistent arousals may have interfered with accuracy of AI 2 [conventional]."* Further, their interaction experience with the same AI assistant can also shape their judgement of where they will likely disagree with the system: *"On 'B' [ambiguity-aware], I tried to focus more on the ambiguous epochs indicated by the AI and then on the staging that the AI in 'A' [conventional] did not perform well with."* AI assistants could leverage this insight by grouping contentious cases based on an expert's reviewing and correction behaviour to adjust to their internal representation of specific types of ambiguity.

DISCUSSION

In this work, we studied how highlighting and explaining ambiguity by AI assistants can aid medical experts in their decision making for contentious clinical cases. We conducted

a within-subjects study to investigate the use of ambiguity-aware AI assistants by medical experts. Our results show that the ambiguity-aware AI can alter experts' workflows by increasing the proportion of contentious cases reviewed while maintaining overall productivity.

While experts' overall labeling accuracy was not affected by providing ambiguity-awareness, we observed a significant effect of argument relevance on experts' case correction rate. This promising insight motivates future research into the development and validation of ambiguity-aware AI systems capable of providing highly relevant ambiguity explanations for previously unseen cases.

Experts' overall preferences and perceived levels of trust for either AI were polarized. Results suggested higher perceived integrity, and a trend towards higher confidence in the ambiguity-aware AI assistant compared to the conventional variant. These mixed results may indicate the existence of other latent variables (e.g., experts' familiarity with or trust in automation technology) which could shape experts' perception of AI systems generally. Here, we discuss the generalizability and design implications of our findings and conclude with limitations of our study and directions for future work.

Design Implications for AI-based CDS Systems

Our findings have implications for different stages in the design of AI-based CDS systems, ranging from data collection over model training to the design of user interfaces for AI systems.

Data collection. In our work, we simulate an AI assistant's capability to identify multiple conflicting arguments for why a medical classification decision may be contentious. To this end, we rely on discussion metadata from a previous study on collective adjudication among medical experts [49]. Developing an AI system capable of generating ambiguity explanations for previously *unseen* cases would require that structured information on contentious cases is given in the training data. While several approaches have been suggested to collect unstructured, open-ended arguments for contentious classification cases [12, 16, 51, 55], recent work from the medical domain demonstrates that imposing structure on the discussion process can facilitate a deeper understanding of expert disagreement [48, 49] and accelerate consensus formation [53]. We recommend that data collection procedures for AI-based CDS systems be equipped with structured discussion procedures to benefit from these findings and facilitate the development of ambiguity-aware classification models.

Model training. Our study suggests that expert workflows and trust can be positively affected by endowing AI-based CDS systems with the ability to not only make classification suggestions, but also to identify which cases may be contentious and why. Implementation of such systems would require that supervised machine learning models are equipped with additional prediction targets beyond classification labels alone. These additional prediction targets could include the likelihood and potential sources of expert disagreement. They could be integrated either into one joint training process or by developing several separate models, one for each target. Cohen et al. [13] describe some additional requirements and challenges in this context.

User interfaces. In this work, we evaluate one specific way of displaying and explaining ambiguity to expert end users by visually emphasizing contentious cases within a collection of cases and by providing text-based arguments for conflicting classification choices. While our results suggest that this representation may be effective, we recommend that future work may explore more complex design considerations such as prioritization of cases based on their disagreement likelihood, and interactive filters to group cases which may be contentious for similar reasons.

Generalizability

Our study sheds light on the use of ambiguity-awareness in the specific domain of sleep stage classification based on biomedical time series data. Therefore, caution is warranted in generalizing the results of this study to outside domains. However, we argue that similar displays of ambiguity explanations can be useful for various types of medical assessments because the issues motivating our study are prevalent across subspecialties.

Despite the abundance of standardized medical guidelines [4], expert disagreement is prevalent across medical disciplines [5, 54], making our approach useful beyond the specific domain of sleep health. For example, differential diagnosis of epilepsy requires that specialized neurologists visually inspect EEG data similar in nature to that used in our study. Ambiguity-aware AI assistants could support the small pool of specialists world-wide in detecting epileptiform abnormalities [3] and thus increase access to healthcare for patients with epilepsy in low- and middle-income countries [64, 65].

The issue of expert disagreement in medical assessments has also been addressed using structured adjudication for other data modalities, e.g., assessment of retinal images for diabetic retinopathy grading [52, 53] or glaucoma risk assessment [24, 39]. These studies suggest that the recommendations we make for data collection in this work have been considered independently and may be of merit beyond the development of ambiguity-aware AI systems.

Limitations and Future Work

In this work, we conducted a within-subjects study to investigate the use of ambiguity-aware AI assistants by medical experts. Due to the tight working schedule of our experts and the remote nature of our study, it was challenging to control the timing of each step in the experiment precisely. For instance, participants varied in how long they waited after completing the first main task before starting the second one. This lack in experimental control may have impacted the extent to which exposure to the first AI assistant affected how experts interacted with the latter one.

In our Wizard-of-Oz study, the ambiguity-aware AI was *simulated*, in the sense that the assistant presented ambiguity information and arguments generated from real expert discussions. While prior work has demonstrated the potential of predicting the likelihood of expert disagreement directly from raw

medical data [42], future work can focus on training machine-learning algorithms based on ambiguity explanation data to provide human-interpretable arguments for previously unseen contentious cases.

Finally, related work shows that medical practitioners seek to understand the specific strengths and weaknesses of an AI *before* interacting with it [9]. Our work offers similar findings by showing that explaining AI uncertainty can be useful also *during* the interaction and help experts allocate cognitive resources and reassess their level of trust appropriately for each specific case. While we did not detect a significant effect of ambiguity tolerance on overall AI preference, we observed a trend that experts with higher ambiguity tolerance exhibited more polarized preferences towards either AI assistant. Future research may explore how different variables such as personality traits [7], domain-specific and culture-specific communication styles [47] may shape these expectations and perceptions on the side of medical experts.

CONCLUSION

In this work, we provided a novel perspective on the problem of how AI assistants for medical reasoning can explain ambiguous cases to human experts. Our results from a user study with twelve medical experts comparing a conventional AI assistant to a simulated ambiguity-aware AI assistant suggest that the system's ability to not only flag, but also explain contentious patient cases has merits for end users. In particular, we observed that in comparison to the conventional AI, the ambiguity-aware AI was more effective in guiding experts' attention to contentious medical cases. In addition, our results demonstrate that if explanations contain irrelevant arguments, experts' accuracy at correcting AI-suggested labels can drop below 50%. Our work has implications for the design of AI-based technology not only in the field of medicine, but more broadly in fields that face similar challenges with classification ambiguity and expert disagreement.

ACKNOWLEDGMENTS

We thank Rui de Sousa for his invaluable help in recruiting participants for this study. This work was funded by NSERC CHRP (CHRP 478468-15), CIHR CHRP (CPG-140200), and the Google PhD Fellowship Program.

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 - CHI '18*. ACM Press, New York, New York, USA, 1–18. DOI: http://dx.doi.org/10.1145/3173574.3174156

[2] Lora Aroyo and Chris Welty. 2014. The three sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34.

[3] Elham Bagheri, Justin Dauwels, Brian C. Dean, Chad G. Waters, M. Brandon Westover, and Jonathan J. Halford. 2017. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clinical Neurophysiology* 128, 10 (10 2017), 1994–2005. DOI: http://dx.doi.org/10.1016/j.clinph.2017.06.252

[4] A. Baker, K. Young, J. Potter, and I. Madan. 2010. A review of grading systems for evidence-based guidelines produced by medical specialties. *Clinical Medicine* 10, 4 (8 2010), 358–363. DOI: http://dx.doi.org/10.7861/clinmedicine.10-4-358

[5] Michael L. Barnett, Dhruv Boddupalli, Shantanu Nundy, and David W. Bates. 2019. Comparative Accuracy of Diagnosis by Collective Intelligence of Multiple Physicians vs Individual Physicians. *JAMA Network Open* 2, 3 (3 2019), e190096. DOI: http://dx.doi.org/10.1001/jamanetworkopen.2019.0096

[6] Brian H. Bornstein and A. Christine Emler. 2001. Rationality in medical decision making: a review of the literature on doctors decision-making biases. *Journal of Evaluation in Clinical Practice* 7, 2 (5 2001), 97–107. DOI: http://dx.doi.org/10.1046/j.1365-2753.2001.00284.x

[7] Stanley Budner. 1962. Intolerance of ambiguity as a personality variable. *Journal of Personality* 30, 1 (3 1962), 29–50. DOI: http://dx.doi.org/10.1111/j.1467-6494.1962.tb02303.x

[8] 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. 2019a. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making. http://arxiv.org/abs/1902.02960

[9] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019b. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–24. DOI:http://dx.doi.org/10.1145/3359206

[10] Arthur Carvalho and Kate Larson. 2013. A Consensual Linear Opinion Pool. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press, Beijing, China, 2518–2524. http://dl.acm.org/citation.cfm?id=2540128.2540491

[11] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM, ACM Press, New York, New York, USA, 2334–2346. DOI: http://dx.doi.org/10.1145/3025453.3026044

[12] Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Daniel S. Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human*

Factors in Computing Systems - CHI '19. ACM Press, New York, New York, USA, 1–14. DOI: http://dx.doi.org/10.1145/3290605.3300761

[13] Robin Cohen, Mike Schaekermann, Sihao Liu, and Michael Cormier. 2019. Trusted AI and the Contribution of Trust Modeling in Multiagent Systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1644–1648. http://dl.acm.org/citation.cfm?id=3306127.3331890

[14] David A. Cook, Jonathan Sherbino, and Steven J. Durning. 2018. Management Reasoning - Beyond the Diagnosis. *JAMA* 319, 22 (6 2018), 2267. DOI: http://dx.doi.org/10.1001/jama.2018.4385

[15] Suzette Cooke and Jean-Francois Lemay. 2017. Transforming Medical Assessment: Integrating Uncertainty Into the Evaluation of Clinical Reasoning in Medical Education. *Academic medicine : journal of the Association of American Medical Colleges* 92, 6 (2017), 746–751. DOI: http://dx.doi.org/10.1097/ACM.0000000000001559

[16] Ryan Drapeau, Lydia B. Chilton, Jonathan Bragg, and Daniel S. Weld. 2016. MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[17] Stephan Dreiseitl and Michael Binder. 2005. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artificial Intelligence in Medicine* 33, 1 (1 2005), 25–30. DOI: http://dx.doi.org/10.1016/j.artmed.2004.07.007

[18] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19*. ACM Press, New York, New York, USA, 263–274. DOI: http://dx.doi.org/10.1145/3301275.3302316

[19] Adrian Galdran, M. Meyer, P. Costa, MendonCa, and A. Campilho. 2019. Uncertainty-Aware Artery/Vein Classification on Retinal Images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 556–560. DOI: http://dx.doi.org/10.1109/ISBI.2019.8759380

[20] Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748.

[21] Cosima Gretton. 2018. Trust and Transparency in Machine Learning-Based Clinical Decision Support. 279–292. DOI: http://dx.doi.org/10.1007/978-3-319-90403-0{_}14

[22] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *AAAI Conference on Artificial Intelligence*. https://arxiv.org/pdf/1703.08774.pdf

[23] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. ACM, ACM Press, New York, New York, USA, 3511–3522. DOI: http://dx.doi.org/10.1145/3025453.3025781

[24] Naama Hammel, Mike Schaekermann, Sonia Phene, Carter Dunn, Lily Peng, Dale R Webster, and Rory Sayres. 2019. A Study of Feature-based Consensus Formation for Glaucoma Risk Assessment. *Investigative Ophthalmology & Visual Science* 60, 9 (2019), 164.

[25] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. 139–183. DOI: http://dx.doi.org/10.1016/S0166-4115(08)62386-9

[26] Hayley K. Jach and Luke D. Smillie. 2019. To fear or fly to the unknown: Tolerance for ambiguity and Big Five personality traits. *Journal of Research in Personality* 79 (4 2019), 67–78. DOI: http://dx.doi.org/10.1016/j.jrp.2019.02.003

[27] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (3 2000), 53–71. DOI: http://dx.doi.org/10.1207/S15327566IJCE0401{_}04

[28] Samed Jukić and Jasmin Kevrić. 2018. Majority Vote of Ensemble Machine Learning Methods for Real-Time Epilepsy Prediction Applied on EEG Pediatric Data. *TEM Journal* 7, 2 (2018), 313.

[29] Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 5092–5103. DOI: http://dx.doi.org/10.1145/2858036.2858558

[30] Gideon Keren and Léonie E.M. Gerritsen. 1999. On the robustness and possible accounts of ambiguity aversion. *Acta Psychologica* 103, 1-2 (11 1999), 149–172. DOI: http://dx.doi.org/10.1016/S0001-6918(99)00034-7

[31] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–14. DOI:http://dx.doi.org/10.1145/3290605.3300641

[32] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* (3 2018). DOI: http://dx.doi.org/10.1016/j.ophtha.2018.01.034

[33] Vera P. Luther and Sonia J. Crandall. 2011. Commentary: Ambiguity and Uncertainty: Neglected Elements of Medical Education Curricula? *Academic Medicine* 86, 7 (7 2011), 799–800. DOI: http://dx.doi.org/10.1097/ACM.0b013e31821da915

[34] V K Chaithanya Manam and Alexander J Quinn. 2018. WingIt: Efficient Refinement of Unclear Task Instructions. In *The Sixth AAAI Conference on Human Computation and Crowdsourcing*. 108–116. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17931

[35] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, New York, New York, USA, 279–288. DOI: http://dx.doi.org/10.1145/3287560.3287574

[36] Geoffrey R. Norman, Lawrence E. M. Grierson, Jonathan Sherbino, Stanley J. Hamstra, Henk G. Schmidt, and Silvia Mamede. 2018. Expertise in Medicine and Surgery. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press, 331–355. DOI: http://dx.doi.org/10.1017/9781316480748.019

[37] J. A. Osheroff, J. M. Teich, B. Middleton, E. B. Steen, A. Wright, and D. E. Detmer. 2007. A Roadmap for National Action on Clinical Decision Support. *Journal of the American Medical Informatics Association* 14, 2 (3 2007), 141–145. DOI: http://dx.doi.org/10.1197/jamia.M2334

[38] Anh T Pham, Raviv Raich, and Xiaoli Z Fern. 2017. Dynamic programming for instance annotation in multi-instance multi-label learning. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2381–2394.

[39] Sonia Phene, R. Carter Dunn, Naama Hammel, Yun Liu, Jonathan Krause, Naho Kitade, Mike Schaekermann, Rory Sayres, Derek J. Wu, Ashish Bora, Christopher Semturs, Anita Misra, Abigail E. Huang, Arielle Spitze, Felipe A. Medeiros, April Y. Maa, Monica Gandhi, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* (9 2019). DOI: http://dx.doi.org/10.1016/j.ophtha.2019.07.024

[40] Stefan Räbiger, Gizem Gezici, Myra Spliliopoulou, and Yücel Sayg\in. 2018. Predicting worker disagreement for more effective crowd labeling. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE.*

[41] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, New York, New York, USA, 1–13. DOI: http://dx.doi.org/10.1145/3173574.3173677

[42] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2018. Direct Uncertainty Prediction for Medical Second Opinions. (7 2018). http://arxiv.org/abs/1807.01771

[43] Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. (7 2017). http://arxiv.org/abs/1707.01836

[44] Paisan Raumviboonsuk, Jonathan Krause, Peranut Chotcomwongse, Rory Sayres, Rajiv Raman, Kasumi Widner, Bilson J. L. Campana, Sonia Phene, Kornwipa Hemarat, Mongkol Tadarati, Sukhum Silpa-Archa, Jirawut Limwattanayingyong, Chetan Rao, Oscar Kuruvilla, Jesse Jung, Jeffrey Tan, Surapong Orprayoon, Chawawat Kangwanwongpaisan, Ramase Sukumalpaiboon, Chainarong Luengchaichawang, Jitumporn Fuangkaew, Pipat Kongsap, Lamyong Chualinpha, Sarawuth Saree, Srirut Kawinpanitan, Korntip Mitvongsa, Siriporn Lawanasakol, Chaiyasit Thepchatri, Lalita Wongpichedchai, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2019. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *npj Digital Medicine* 2, 1 (12 2019), 25. DOI: http://dx.doi.org/10.1038/s41746-019-0099-8

[45] D A Redelmeier and E Shafir. 1995. Medical decision making in situations that offer multiple alternatives. *JAMA* 273, 4 (1 1995), 302–5. http://www.ncbi.nlm.nih.gov/pubmed/7815657

[46] Richard S. Rosenberg and Steven van Hout. 2013. The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring. *Journal of Clinical Sleep Medicine* (1 2013). DOI: http://dx.doi.org/10.5664/jcsm.2350

[47] Elaheh Sanoubari, Stela H. Seo, Diljot Garcha, James E. Young, and Veronica Loureiro-Rodriguez. 2019. Good Robot Design or Machiavellian? An In-the-Wild Robot Leveraging Minimal Knowledge of Passersby's Culture. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 382–391. DOI: http://dx.doi.org/10.1109/HRI.2019.8673326

[48] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019a. Capturing Expert Arguments from Medical

Adjudication Discussions in a Machine-readable Format. In *Companion Proceedings of The 2019 World Wide Web Conference - WWW '19*, Vol. 2. ACM Press, New York, New York, USA, 1131–1137. DOI: http://dx.doi.org/10.1145/3308560.3317085

[49] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019b. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. In *Proceedings of the 2019 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2019)*, Vol. 3. Austin, TX, 1–23. DOI: http://dx.doi.org/10.1145/3359178

[50] Mike Schaekermann, Carrie J Cai, Abigail E Huang, and Rory Sayres. 2020. Expert Discussions Improve Comprehension of Difficult Cases in Medical Image Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems - CHI '20*. ACM Press, Honolulu, HI, USA. DOI: http://dx.doi.org/10.1145/3313831.3376290

[51] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. In *Proceedings of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2018)*, Vol. 2. New York City, NY, 1–19. DOI: http://dx.doi.org/10.1145/3274423

[52] Mike Schaekermann, Naama Hammel, Brian Basham, Bilson Campana, Edith Law, Lily Peng, Dale R Webster, and Rory Sayres. 2019a. Asynchronous Remote Adjudication for Grading Diabetic Retinopathy. *Investigative Ophthalmology & Visual Science* 60, 9 (2019), 158.

[53] Mike Schaekermann, Naama Hammel, Michael Terry, Tayyeba K. Ali, Yun Liu, Brian Basham, Bilson Campana, William Chen, Xiang Ji, Jonathan Krause, Greg S. Corrado, Lily Peng, Dale R. Webster, Edith Law, and Rory Sayres. 2019b. Remote Tool-Based Adjudication for Grading Diabetic Retinopathy. *Translational Vision Science & Technology* 8, 6 (12 2019), 40. DOI:http://dx.doi.org/10.1167/tvst.8.6.40

[54] Mike Schaekermann, Edith Law, Kate Larson, and Andrew Lim. 2018. Expert Disagreement in Sequential Labeling: A Case Study on Adjudication in Medical Time Series Analysis. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing at HCOMP 2018*. Zurich, Switzerland.

[55] Mike Schaekermann, Edith Law, Alex C Williams, and William Callaghan. 2016. Resolvable vs. Irresolvable Ambiguity: A New Hybrid Framework for Dealing with Uncertain Ground Truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI 2016*. San Jose, CA.

[56] Hanna Schneider, Julia Wayrauther, Mariam Hassib, and Andreas Butz. 2019. Communicating Uncertainty in Fertility Prognosis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–11. DOI:http://dx.doi.org/10.1145/3290605.3300391

[57] Jens B. Stephansen, Alexander N. Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, Yun L. Sun, Yves Dauvilliers, Sabine Scholz, Lucie Barateau, Birgit Hogl, Ambra Stefani, Seung Chul Hong, Tae Won Kim, Fabio Pizza, Giuseppe Plazzi, Stefano Vandi, Elena Antelmi, Dimitri Perrin, Samuel T. Kuna, Paula K. Schweitzer, Clete Kushida, Paul E. Peppard, Helge B. D. Sorensen, Poul Jennum, and Emmanuel Mignot. 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications* 9, 1 (12 2018), 5229. DOI: http://dx.doi.org/10.1038/s41467-018-07229-3

[58] Evangelia Tsiga, Efharis Panagopoulou, Nick Sevdalis, Anthony Montgomery, and Alexios Benos. 2013. The influence of time pressure on adherence to guidelines in primary care: an experimental study. *BMJ Open* 3, 4 (4 2013), e002700. DOI: http://dx.doi.org/10.1136/bmjopen-2013-002700

[59] A. Tversky and D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (9 1974), 1124–1131. DOI: http://dx.doi.org/10.1126/science.185.4157.1124

[60] Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. 2019. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science* 6, 5 (5 2019), 181870. DOI:http://dx.doi.org/10.1098/rsos.181870

[61] Michael Van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 900–907.

[62] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, New York, New York, USA, 1–15. DOI:http://dx.doi.org/10.1145/3290605.3300831

[63] Simon C Warby, Sabrina L Wendt, Peter Welinder, Emil G S Munk, Oscar Carrillo, Helge B D Sorensen, Poul Jennum, Paul E Peppard, Pietro Perona, and Emmanuel Mignot. 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. *Nature Methods* 11, 4 (2 2014), 385–392. DOI: http://dx.doi.org/10.1038/nmeth.2855

[64] Jennifer Williams, Fodé Abass Cisse, Mike Schaekermann, Foksuna Sakadi, Nana Rahamatou Tassiou, Aissatou Kenda BAH, Abdoul Bachir Djibo Hamani, Andrew Lim, Edward C W Leung, Tadeu A Fantaneau, Tracey Milligan, Vidita Khatri, Daniel Hoch, Manav Vyas, Alice Lam, Gladia Hotan, Joseph Cohen, Edith Law, and Farrah Mateen. 2019a. Utilizing a wearable smartphone-based EEG for pediatric epilepsy patients in the resource poor environment of Guinea: A prospective study. *Neurology* 92, 15 Supplement (2019). https://n.neurology.org/content/92/15_Supplement/N5.001

[65] Jennifer A Williams, Fodé Abass Cisse, Mike Schaekermann, Foksouna Sakadi, Nana Rahamatou Tassiou, Gladia C. Hotan, Aissatou Kenda Bah, Abdoul Bachir Djibo Hamani, Andrew Lim, Edward C.W. Leung, Tadeu A. Fantaneanu, Tracey A. Milligan, Vidita Khatri, Daniel B. Hoch, Manav V. Vyas, Alice D. Lam, Joseph M. Cohen, Andre C. Vogel, Edith Law, and Farrah J. Mateen. 2019b. Smartphone EEG and remote online interpretation for children with epilepsy in the Republic of Guinea: Quality, characteristics, and practice implications. *Seizure* 71 (10 2019), 93–99. DOI:http://dx.doi.org/10.1016/j.seizure.2019.05.025