# Human-Centered Deferred Inference: Measuring User Interactions and Setting Deferral Criteria for Human-AI Teams

Stephan J. Lemmer
University of Michigan
Ann Arbor, MI, USA
lemmersj@umich.edu

Anhong Guo
University of Michigan
Ann Arbor, MI, USA
anhong@umich.edu

Jason J. Corso
University of Michigan
Ann Arbor, MI, USA
jjcorso@umich.edu

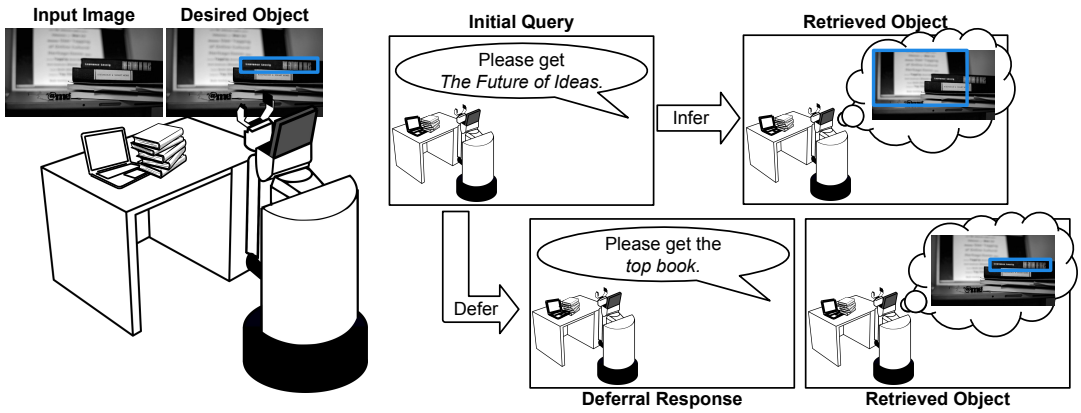


**Figure 1: Although deep learning provides high performance in some settings, the constraint of a one-time human input combined with counterintuitive gaps in knowledge can lead to poor performance in interaction scenarios. For example: if a robot asked to retrieve a book by its title (*please get The Future of Ideas*) is forced to infer it will identify the wrong object, but if it is allowed to defer—soliciting additional information such as *please get the top book* from the human—it will behave correctly. Images and expressions from RefCOCO [39], predictions are from UNITER [18].**

## ABSTRACT

Although deep learning holds the promise of novel and impactful interfaces, realizing such promise in practice remains a challenge: since dataset-driven deep-learned models assume a one-time human input, there is no recourse when they do not understand the input provided by the user. Works that address this via deferred inference—soliciting additional human input when uncertain—show meaningful improvement, but ignore key aspects of how users and models interact. In this work, we focus on the role of users in deferred inference and argue that the deferral criteria should be a function of the user and model as a team, not simply the model itself. In support of this, we introduce a novel mathematical formulation, validate it via an experiment analyzing the interactions of 25 individuals with a deep learning-based visiolinguistic model, and identify user-specific dependencies that are under-explored in prior work. We conclude by demonstrating two human-centered procedures for setting deferral criteria that are simple to implement, applicable to a wide variety of tasks, and perform equal to or better than equivalent procedures that use much larger datasets.


*IUI '23, March 27–31, 2023, Sydney, NSW, Australia*

ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584092

## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Natural language interfaces**; **Empirical studies in HCI**.

## KEYWORDS

deferred inference, neural networks, referring expression comprehension

**ACM Reference Format:**
Stephan J. Lemmer, Anhong Guo, and Jason J. Corso. 2023. Human-Centered Deferred Inference: Measuring User Interactions and Setting Deferral Criteria for Human-AI Teams. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3581641.3584092

## 1 INTRODUCTION

Deep learning holds the promise of novel interfaces, many of which could have significant practical impact: visual question answering models [1] are being studied as a method to help visually impaired individuals understand the visual world [7, 8, 31], referring expression comprehension [56] is a critical technology for robots in, for example, an elder-care setting [13, 40, 59, 84], vision-and-dialog navigation [78] will simplify control of search and rescue vehicles [5, 10], among others [53, 66, 83]. Despite these human-centered motivations, the formulation of supervised deep learning—a model is given an input and rewarded for a correct output—means that

there is no recourse when the human provides information that is semantically ambiguous [6, 30] or mismatched with the features learned by the model [77].

For illustration, consider the case shown in Figure 1, where the goal is for the support robot to retrieve a specific book from the desk. The initial query refers to the book by its title, *The Future of Ideas*. Although the robot is unable to correctly resolve the object using this query, standard deep learning formulations would force the robot to *infer*, leading to retrieval of the laptop instead of the target book. Because of the latency of this process—the human must wait for the robot to retrieve the desired object—it is impractical to evaluate the received object and reformulate the query if the model's output is incorrect as is common practice in search or conversational virtual assistants [71]. Other settings have similar limitations: vision-and-dialog navigation [78] also has significant latency after inference, while visual question answering for the visually impaired [31] does not have a simple method for confirming the model's output.

Recognizing this, some works propose methods for *deferred inference*: when the model is uncertain, it can instead *defer* (Figure 1-bottom) and request additional information in a way that does not require the human to perform the task in place of the AI agent. Approaches to deferred inference include generating follow-up questions [57, 62, 74], using natural language to revise plans [73], and asking for a rephrase [34, 47]. Although such works have demonstrated that deferred inference can be used to reduce error, they often downplay the role of the individual with whom the AI is interacting: Lemmer *et al.* [47] provide a comprehensive evaluation of deferred inference in aggregate but do not consider *deferral criteria*—the exact conditions that result in deferral—while other works [34, 57, 62, 74] select their deferral criteria based on pre-defined properties of the model's output (*e.g.,* margin [34, 57]) without considering their effect on properties such as error or deferral rate, or qualities of an individual user. In this work, we focus on how the choice of deferral criteria must explicitly consider the interaction between an individual and a deep-learned model.

We begin by describing a novel formulation for setting deferral criteria that explicitly considers the individual, the model, and the goals of deferred inference. We validate this formulation via a study with 25 participants on a language-based image cropping task—the same technology underlying the example in Figure 1. Specifically, we identified four major findings: *(i)* there exists a significant relationship between user satisfaction and both error and deferral rate, motivating deliberate setting of deferral criteria; *(ii)* the distribution of output confidences is dependent on the individual, reinforcing the need for user-specific deferral criteria; *(iii)* the *deferral response*—information provided by the user after deferral—is less meaningful to the model than the initial query, demonstrating a shortcoming of reformulation approaches; and *(iv)* the relationship between the model's confidence and error is most likely to be independent of the user, shortening the calibration process when the goal is to target an error. We then demonstrate two methods for setting deferral criteria based on individual users, and find that they perform as well or better than using large datasets, despite having two orders of magnitude fewer calibration examples.

# 2 RELATED WORK

## 2.1 User Interaction with Deep Learned Models

Many applications use a human input to define the task or provide additional information to improve performance: visual question answering [1] requires a human to ask a question and provide an image, keypoint-conditioned viewpoint estimation [76] allows a human to provide image semantics, voice-based video navigation [17] allows a human to provide verbal cues to navigate a video, among others [53, 56, 66, 78, 83]. Because the structure of deep neural networks—a single input produces a single output—solutions to such problems either evaluate error via independent inferences on a dataset [1, 31, 56] or consider team performance from perspectives such as trust and explainability [4, 14, 49, 67, 72], introducing novel interfaces [12, 37, 44], performing satisfaction surveys [21, 52, 53, 86], or evaluating how users respond in a failure case [33, 71].

Other works that explore humans teamed with deep-learned models ignore the qualities of the model and use simplifying assumptions such as Wizard-of-Oz studies [69, 88], simplified computational models [2, 16, 17], or only identifying inputs with incorrect or insufficient semantics [6, 54, 64]. While these areas of research are meaningful, the often counter-intuitive nature of deep learning models [70, 75] means that human input being semantically correct is neither necessary nor sufficient to produce the correct answer [46, 48]. In this work, we compensate for these shortcomings by evaluating human interaction with the deep-learned model in the loop.

## 2.2 Conditional Inference

It is intuitive for a model—human-in-the-loop or not—to only make a final decision when it is confident. In some cases this is done by using the AI to assist decision making by providing relevant information, such as proposed outputs [38, 80], or a prediction with an explanation or confidence value [4, 60, 87]. In other cases, framed as *selective prediction* [19, 20, 23, 27, 28, 43, 45, 81], the AI sends low-confidence inferences to a human—a "second opinion" in medical terms [9, 41, 65]. While these approaches are useful in many cases, they require a human to perform the inference itself when the model is not confident, which is impractical in important use cases such as visual question answering for accessibility [31] or giving verbal commands to a household robot [84]. Because of this, it is important to consider methods for conditional inference that do not require the human to operate in the output space (*e.g.,* fetching the desired object or answering visual sub-questions).

A handful of works, sometimes grouped under the umbrella of *deferred inference* [47] do this by asserting that the initial and subsequent human inputs are made in the same space. Such works generally take intuitive approaches such as generating complementary text queries [57, 62, 74, 79], asking for a rephrase [34, 47], or allowing the human to identify and resolve local minima in tasks with a long time horizon (*e.g.,* adding instructions to a pick-and-place task) [73]). While such approaches are intuitive, they have two shortcomings: first, they often require novel approaches to produce meaningful follow-up questions. Due to the opaque nature of deep neural networks, this will often require assumptions on the task (*e.g.,* iterating through identified objects [57]) or datasets that

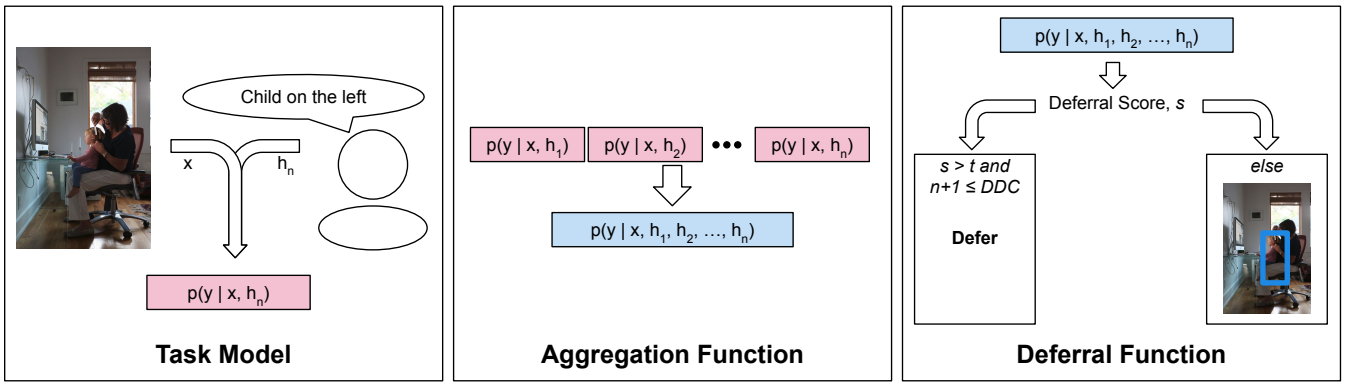


**Figure 2: Deferred inference can be abstracted into three components: the *task model* uses human information and fixed inputs to predict a target value, the *aggregation function* combines multiple outputs from the task model, and the *deferral function* determines whether to perform the inference or defer based on the deferral criteria (is $s > t$?) and whether the deferral depth constraint has been reached. The application of Referring Expression Comprehension [56] is used for illustration.**

lead to trivial follow-up questions (*e.g.,* asking *is there a tined utensil to the left of the pizza?* as a follow-up question to *on which side of the plate is the fork?* [79]). Second, user studies in such works do not typically analyze the interaction between individual users and the deep-learned model, instead setting deferral criteria a-priori and reporting the change in success rate over the deferral-free condition. To address the former shortcoming, we use the approach of Lemmer & Corso [47], while for the latter we perform a user-centered study with the explicit goal of setting deferral criteria.

## 3 BACKGROUND AND RESEARCH QUESTIONS

Throughout this work, we seek a method for setting deferral criteria that meaningfully improves the user's satisfaction with the interaction, which we show is dependent on the error and the deferral rate. Such a method requires not only an understanding of how the model responds to a human input, but also how the input—and the model's response to it—varies from user to user. We begin by discussing the framework and terminology of deferred inference that we use, then provide a theoretical formulation for calculating both the deferral rate and the overall error. These formulations provide important guidance to the questions that must be answered to appropriately set deferral criteria.

### 3.1 Deferred Inference

Deferred inference [47] improves the performance of a human-AI team by allowing the AI to defer—request additional information from the human—when some set of conditions are met. Throughout this work, we formulate deferred inference as the interaction between the three components shown in Figure 2. The first component is the *task model*, $f(x, h_n)$, which uses a fixed input (*e.g.,* an image), $x$, and a human-provided input (*e.g.,* a text query), $h_n$, to produce a distribution across outputs, $p(y|x, h_n)$. There is no hard restriction on what this distribution is: it could be, for example, a softmax across a set of answers in visual [1] or text [66] question answering, a distribution across locations in visual object tracking [22], or a variety of other task-dependent outputs. The second component is the *Aggregation Function*, $h(x, h_1, ..., h_n|f)$, which produces a new belief, $p(y|x, h_1, ..., h_n)$, by combining multiple outputs from the task model. This may be done by any number of methods, such as direct replacement [48], finding the mean distribution [34], or performing a belief update [47]. Throughout this work, we use the last method as our aggregation function and allow the model to defer by asking the user to try again. This approach has the benefit of allowing us to rapidly implement deferred inference on new architectures without needing to develop corresponding text generation architectures [57] or relevant datasets [79].

The output of the aggregation function is passed to the *deferral function*, $g(x, h_1, ..., h_n) \in \{0, 1\}$, which determines whether or not inference should be deferred based on whether some *deferral criteria* has been met. Setting the deferral criteria is the main motivation of this work: it is typically a threshold that is applied to a continuous *deferral score*, such as entropy [47] or margin [34, 57], alongside a *Deferral Depth Constraint* (DDC) that limits the number of times the AI is allowed to defer. Such criteria would be set to target an error or *Deferral Rate* (DR), the average number of deferrals per task. Although previous works minimize the role of the deferral criteria by choosing to evaluate at all DRs and DDCs [47, 48] or ignoring the user burden [34, 57, 62, 73, 74, 79], we find that the deferral criteria must be considered, since user satisfaction is directly related to both error and DR.

### 3.2 Theory on Thresholds

When we set deferral criteria, we seek to balance error and user burden. Previous work [34, 57] has downplayed that tradeoff by making the assumption that it is sufficient to set deferral criteria based on characteristics of the task model. In this section, we show that it is impossible to set a deferral criteria that targets an error or DR without explicitly considering the user for whom that criteria is being set. Throughout this work, we set the DDC to one across all evaluations.

*Expected Deferral Rate.* We begin by showing how to calculate the expected DR, $\mathbb{E}(\text{DR}|t,u)$. A deferral occurs if we have a user, $u$, that user produces a score, $s_1$, and that score is greater than the threshold, $t$. This gives us the formula:

$$\mathbb{E}(\text{DR}|t,u) = \int_{s_1} p(s_1 > t, s_1, u) ds_1. \tag{1}$$

If we expand by chain rule, assume the user is given ($p(u) = 1$) and represent $p(s_1 > t)$ with an indicator function, we get:

$$\mathbb{E}(\text{DR}|t,u) = \int_{s_1} \mathbb{1}(s_1 > t) p(s_1|u) ds_1. \tag{2}$$

This demonstrates clearly that while previous work often sets deferral criteria in a user-agnostic way [34, 57], *we cannot target a deferral rate in a user-agnostic manner if $p(s_1)$ is dependent on the user.* This motivates the research question: *do deferral scores differ meaningfully between users?*

*Probability of Error.* To find the probability of error $p(e|t,u)$, we evaluate separately the contribution to error when a deferral does and doesn't occur. When no deferral occurs, we are looking for the condition where the user, $u$, produces a score, $s_1$, that is less than or equal to the threshold, $t$, and there is an error, $e$. Written mathematically:

$$p(e|t,u,s_1 \leq t) = \int_{s_1} p(e, s_1, s_1 \leq t, u) ds_1. \tag{3}$$

The formulation is similar if deferral has occurred, with the addition of the deferral score after the second human input:

$$p(e|t,u,s_1 > t) = \int_{s_2} \int_{s_1} p(e, s_2, s_1, s_1 > t, u) ds_1 ds_2. \tag{4}$$

Since these two conditions are mutually exclusive ($s_1$ is never simultaneously greater than and less than $t$), we can simply sum these two components. If we invoke the same assumptions as in Equation 2, we get:

$$p(e|t,u) = \int_{s_2} \int_{s_1} (p(e|s_2,u)p(s_2|s_1,u)p(s_1|u)\mathbb{1}(s_1 > t) + p(e|s_1,u)p(s_1|u)\mathbb{1}(s_1 \leq t)) ds_1 ds_2. \tag{5}$$

As when targeting a deferral rate, it is critical to consider the relationship between the deferral score, $s_1$, and the user. Additionally, we note two other questions that should be evaluated: first, if the task model's responses to the first and second human inputs are identical, we can find $p(s_2|s_1,u)$ using only initial responses ($p(s_1|u)$), significantly reducing calibration time. In other words, we ask *how do users respond when an inference is deferred?* Second, although works in calibration [29, 51, 58, 82] show a relationship between probability of error and some deferral scores, such works have never considered the role of individual users. If the relationship between probability of error and deferral score is dependent on the individual, we must consider this when finding $p(e|s_1,u)$ and $p(e|s_2,u)$, instead of simply using large datasets. In other words, we ask *does knowing the user provide additional information about the mapping between probability of error and deferral score?*

Explicitly, this leads us to five research questions:

RQ1 *How is user satisfaction related to error and deferral rate?* It is only necessary to pursue a specific error or deferral rate—which requires user-specific deferral criteria—if these factors have an effect on overall satisfaction.

RQ2 *What are the time dependencies of error, e, and deferral score, s?* The lack of a time variable in the above formulations implicitly assumes static distributions. However, previous work [15, 68], as well as common sense assert that the users require some time to develop their mental model. Thresholds should only be set after this mental model has converged.

RQ3 *Do deferral scores differ meaningfully between users?* By not providing user identities, dataset-focused work in deferred inference [34, 47] implicitly assumes that users are interchangeable, while works that evaluate via human experiments [57, 74] set deferral criteria a-priori and do not consider qualities of the individual. If the deferral score is different between users, the deferral criteria will need to be calibrated for individuals.

RQ4 *How do users respond when an inference has been deferred?* Previous work using our chosen deferral formulation has either accepted deferral responses as is—not comparing qualities of the deferral response to the initial query—or broken time dependency entirely through the use of datasets [34, 47]. If the deferral response is significantly different from the initial response, this dependency should be considered in future work. If not, dataset-like approaches could be used to set deferral criteria for higher deferral depth constraints without collecting many deferral responses for each user.

RQ5 *Does knowing the user provide additional information about the mapping between probability of error and deferral score?* Works in model calibration [29, 51, 58, 82] demonstrate a mapping between deferral scores and probability of error, but do not explore if such mapping is consistent across users. If this mapping is not user-dependent, we can construct both $p(e|s_1,u)$ and $p(e|s_2,u)$ prior to interaction with the individual based on large non-user-specific datasets. This would greatly reduce the time required to set deferral criteria.

# 4 EXPERIMENTAL SETUP

## 4.1 Motivating Application

We used referring expression comprehension [56] as our motivating application. In referring expression comprehension (shown in Figures 1 and 2), the user provides a text query that identifies a specific object in an image. The task model accepts both the image and the text query and attempts to identify the object described in the text, either through a bounding box or per-pixel segmentation. We presented this application to our participants as a language-based image cropping task, which was chosen for two reasons: first, cropping is a commonly performed and easily explainable task, meaning little additional instruction was necessary. Second, unlike other embodiments of referring expression comprehension—such as pick-and-place [57]—cropping can be credibly applied to existing datasets (*i.e.,* MSCOCO [50]) and therefore does not require additional model training or dataset procurement.

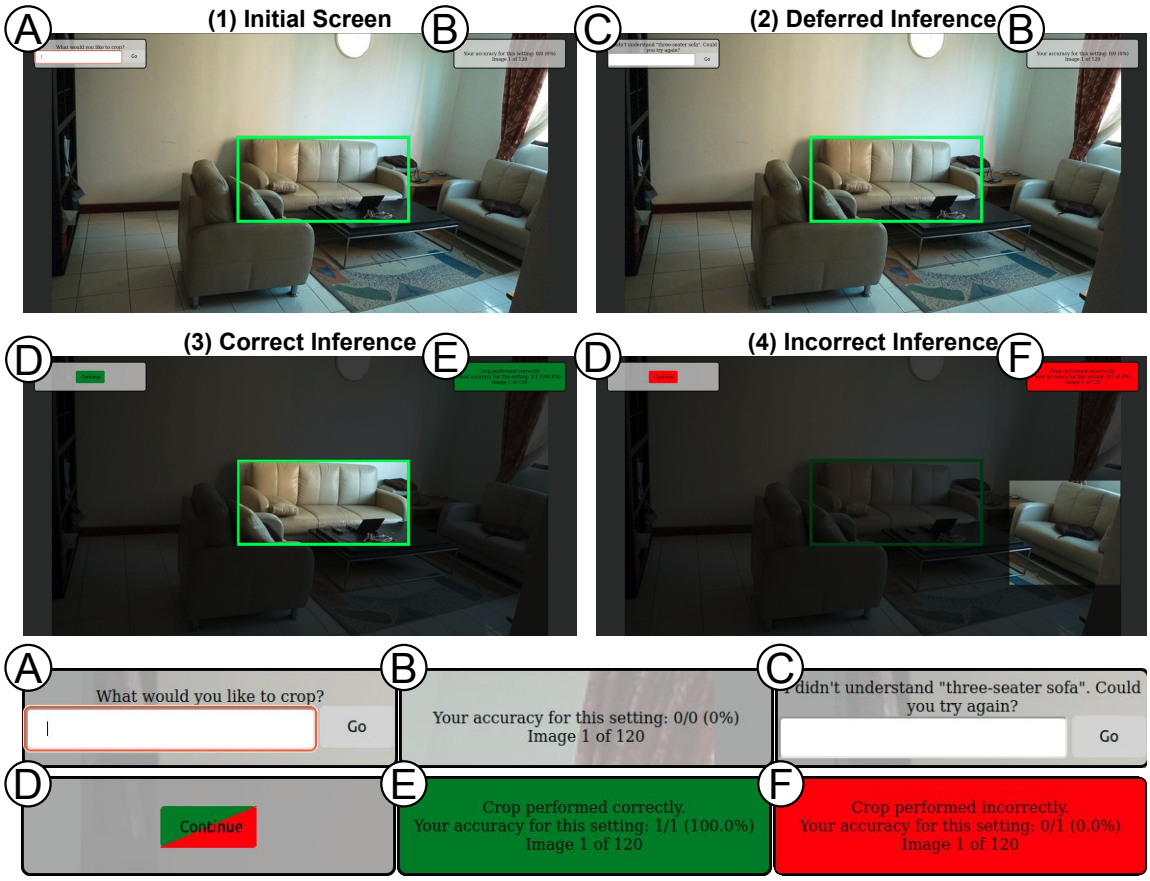


**Figure 3: The four screens in our interface. The user begins in the *Initial Screen* and is tasked with cropping the object in the green box. After entering text on the initial screen, the AI may choose to defer or infer. If the AI chooses to defer, the user is asked to provide another input on the *Deferred Inference* screen. After inference, the user is presented with either the *Correct Inference* screen or the *Incorrect Inference* screen. In both cases, the removed region is darkened. Indicated regions shown below images. The color of region (D) depends on whether the inference was correct. Inputs for correct and incorrect inferences were *three-seater sofa* and *far right sofa*, which were provided by participants to identify the cropped objects.**

As our dataset, we used a subset of target objects from the RefCOCO dataset [39]. This subset was chosen to mitigate two issues observed in our initial tests: first, there were many cases where the target object was visually ambiguous due to a high degree of overlap with other objects in the image—for example, a person standing in front of another. Second, similar to findings on the VQA application [11], there were numerous instances where the model largely ignored the text. Since our focus is on the effect of human input given a clear intent, we selected a subset of RefCOCO that meets the following criteria:

- The object does not have an Intersection-over-Union (IoU) of greater than 0.5 with any other object in the image.
- Of the referring expressions in the RefCOCO dataset [39] for this object, greater than 32% but less than 68% result in a correct answer.

We additionally iterated through the remaining examples to manually remove images that do not clearly indicate a single object or may be offensive, resulting in a total of 1,107 potential crop targets across 842 images. During evaluation, crop targets were randomly picked and an individual participant never saw the same image more than once.

## 4.2 Procedure

*Participants.* We conducted this experiment with 28 adults (older than 18). All participants were required to have normal or corrected-to-normal full-color vision and described themselves as proficient in English. Participants were solicited via local mailing lists and located in the United States at the time of the study. Participants were asked to use a computer with a mouse and keyboard, and were supervised virtually during the experiment. Three participants were identified as malicious or inattentive actors (error greater than three standard deviations above the mean) and their data was excluded from further analysis.

Of the remaining 25 participants, 12 identified as male, 12 identified as female, and 1 preferred not to state. Mean age was 25.2 ± 2.88, technical competence was reported as 5.76 ± 1.24 out of 7,

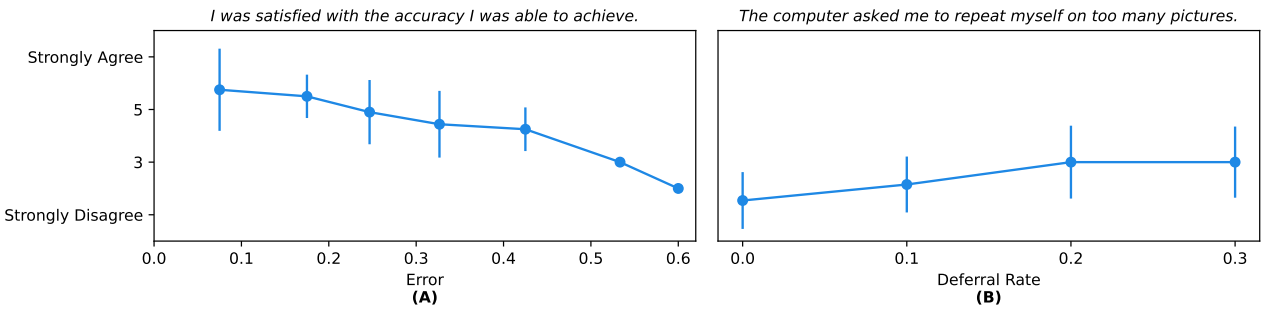


**Figure 4: Relationship between performance measures—error (A) and deferral rate (B)—and user satisfaction. Error is binned at intervals of 10%, and points are plotted at the mean within each bin. Error bars represent one standard deviation.**

and experience with conversational virtual assistants was reported as 4.16 ± 1.76 out of 7. Our study was approved by our institution's IRB, and participants consented to participate in the study before the study started. Participants were compensated $20 for their participation.

*Instructions.* After agreeing to the consent form but prior to any interaction with the system, participants were given a set of instructions for the study. These instructions described the overall goal of image cropping, the interface they would use, the actions the system may take (deferral, correct answer, incorrect answer), and the surveys they would be given. Instructions did not contain any example phrases in order to avoid biasing the user.

*Background Survey.* Participants were asked to provide demographic data (age and gender) as well as their perceived technical competence (1-7 agreement with *I consider myself to be technically adept*), experience with voice assistants in general (1-7 agreement with *I am experienced with voice assistants (Alexa, Siri, etc.)*), and experience with the commercially available voice assistants Amazon Alexa, Apple Siri, Google Assistant, Microsoft Cortana, and Samsung Bixby (*How often do you use the following voice assistants: several times a day/several times a week/1-2 times a week/less*).

*Treatments.* Once participants completed the background survey, they were given four treatments corresponding to deferral rates of 0.0, 0.1, 0.2, and 0.3. The 0.0 deferral rate setting was given first to allow the user to gain familiarity with the system without the noise of random deferrals, while the other three deferral rate settings were presented in a randomized order. Prior to each setting, participants were informed of the beginning a new setting, but no information was provided about which variable was changed.

Participants then interacted with the cropping model via the interface shown in Figure 3. For every task—30 in total for each treatment—they were given a random, previously unseen image with a green box drawn around the target object with the question "what would you like to crop?" (*Initial Screen*). After giving a referring expression corresponding to the boxed object, the model could defer or perform the inference. If the model chose to defer, participants were presented with the last entered phrase, a prompt stating *I didn't understand "[entered text]". Could you try again?* (*Deferred Inference*). If the model chose to perform the inference, the identified object was indicated by shading the removed region. If the crop was correct, the screen showed green (*Correct Inference*), while an incorrect crop showed red (*Incorrect Inference*).[1] The accuracy (number correct, number attempted, and those value expressed as a percent) for the current setting, as well as the overall progress (number of crops performed and total number of crops), were shown on the upper-right corner.

After each treatment, participants were asked to report their satisfaction by rating their agreement with the following statements on a 1-7 Likert scale, where 1 is strongly disagree, and 7 is strongly agree:

- I was satisfied with the accuracy I was able to achieve.
- The computer asked me to repeat myself on too many pictures.

## 4.3 Technical Details

*Task Model.* We used the UNITER architecture [18] as our task model. When used for the application of referring expression comprehension, this model accepts a set of detected objects—we use ground-truth detections to minimize the influence of the object detector—and word embeddings, then outputs a softmax distribution with one value for each input object. As in previous work [47], we train on the RefCOCO dataset [39], and perform Monte Carlo Dropout [24] with 50 forward passes.

*Aggregation Function.* Since the output of the UNITER architecture is a softmax, integrating human information across timesteps is straightforward. We assume inputs are independent, then perform an update on the probability. Written mathematically for $o$ detected objects, this is:

$$p(y_i|x, h_1, ..., h_n) = \frac{\prod_{k=1}^{n} p(y_i|x, h_k)}{\sum_{j=1}^{o} \prod_{k=1}^{n} p(y_j|x, h_k)}, \tag{6}$$

*Deferral Functions.* Throughout this work, we use two different deferral functions. When interacting with our test participants, we seek to precisely target a deferral rate. Since we cannot do this

[1] Due to the common use of color as an attribute in the training dataset [39], we chose to restrict our study to individuals with full-color vision. Under this constraint, we used the red and green color scheme.

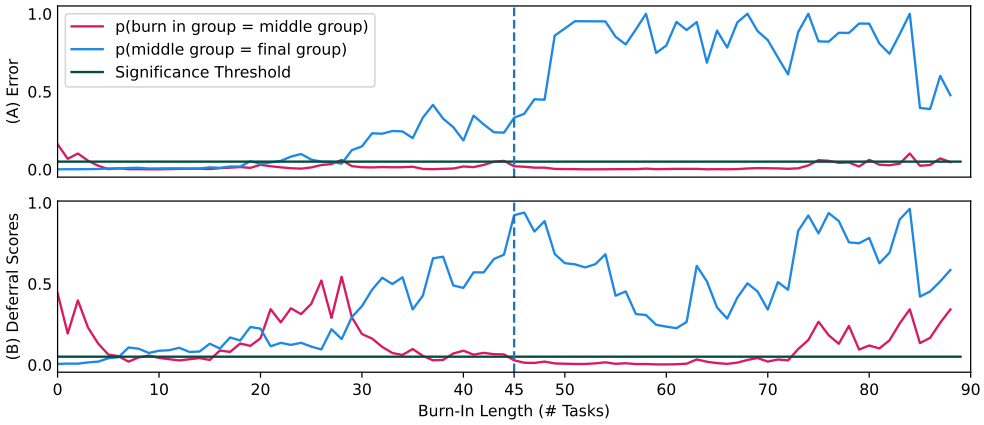


**Figure 5: The probability of the first $n$ samples being different from the next half of the remaining samples (pink line) and the two halves of the remaining samples after $n$ being different (blue line). When the first condition is true and the second is false, represented by the blue vertical line, the users' mental models have settled.**

without establishing deferral criteria—a primary goal for this work—we instead defer randomly such that the exact target deferral rate is reached for every treatment:

$$p(\text{deferral}) = \max(DR, \frac{d_r - d_e}{t_t - t_p})\mathbb{1}(d_e < d_r), \tag{7}$$

where $d_r$ is the number of deferrals required for the target deferral rate (deferral rate times number of tasks in the treatment length), $d_e$ is the number of deferrals that have been executed in this treatment, $t_t$ is the number of tasks in the treatment (30, in our experiments) and $t_p$ is the number of tasks that have been performed.

During our analysis, we use the entropy of the output distribution as our deferral score, matching [47]. This is calculated as:

$$s = -\sum_{j=1}^{o} p(y_j|x, h_1, ..., h_n)\log(p(y_j|x, h_1, ..., h_n)) \tag{8}$$

## 5 RESULTS

### RQ1: Is user satisfaction related to error and deferral rate?

One motivation for our investigation is the assumption that error and/or deferral rate strongly influence user satisfaction. We plot the Likert responses of our treatment surveys against the error (A) and deferral rate (B) in Figure 4. We found that *satisfaction is related to both error and deferral rate*: the lower the error and fewer deferrals, the higher the reported satisfaction, suggesting we can increase user satisfaction by optimally controlling these two variables. For both performance measures, there appears to be a plateau: for error, the satisfaction for error rates between 0 and 10% (7.5% mean) was not significantly different than the satisfaction for error rates between 10% and 20% (17.5% mean) (Mann-Whitney U, $p > 0.10$), but both had a weak significance ($p < 0.10$) compared to an the 20% to 30% range (24.7% mean), and were perceived as better to a significant degree ($p < 0.05$) than the next two bins (32.7% and 42.5% on the mean). Significance could not be established for higher error rates due to small sample sizes. For deferral rate, the results were similar: satisfaction with deferral rate differed significantly (Mann-Whitney U test, $p < 0.05$) between deferral rates of 0.0, 0.1, and 0.2, but deferral rates of 0.2 and 0.3 both reported a mean response of 3 to the question *the computer asked me to repeat myself on too many pictures*. Because satisfaction is related to error and deferral rate, *the ideal approach for deferral is not to set deferral criteria based on model-centric qualities such as margin [34, 57], but to target a deferral rate or error directly using the formulation described in Section 3.2.*

### RQ2: What are the time dependencies of error, $e$, and deferral score, $s$?

In order to accurately set deferral criteria, we must be confident that the distributions we are working with are not changing during the calibration period. If they are—as is likely [2, 3, 15]—any deferral criteria we produce will quickly become inaccurate.

To determine if and when our distributions have settled, we divided the initial queries from all participants into three groups:

(1) The *burn-in* group consists of the first $n$ tasks, where we assume the user is still learning how to interact with the model.
(2) The middle group is the first half of the remaining data.
(3) The final group is the second half of the remaining data.

We considered the burn-in period to be over when 1) the burn-in group is significantly different from the middle group, and 2) the middle group is not significantly different from the final group. If either of these conditions are not true, it indicates either that a substantial number of burn-in scores are within the middle group—$n$ is too small—or a substantial number of settled scores are within the burn-in group—$n$ is too large. Using a Fisher Exact test to measure similarity of error distributions and a Mann-Whitney U test to measure the similarity of deferral score distributions, we found that if we regard $p < 0.05$ as significant, the earliest that both of these conditions are met for several consecutive timesteps was at 45 tasks. For this reason, we use 45 as a burn-in period throughout this work. This is represented visually in Figure 5. We additionally

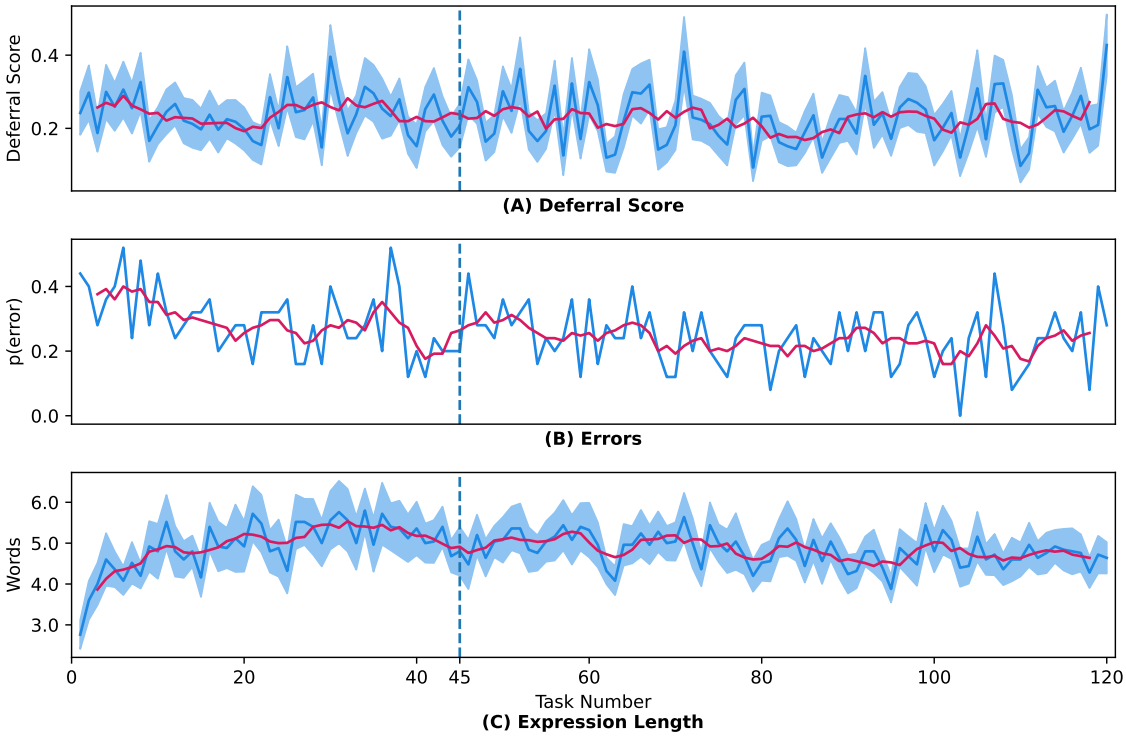


**Figure 6: The relationship between task number and deferral score, probability of error, and expression length. Mean across the five adjacent task numbers shown in pink. Burn-in period shown with a dashed vertical line. Shaded area is one standard error.**

show the mean values for deferral score, errors, and expression length against time in Figure 6, with this burn-in period indicated.

## RQ3: Do deferral scores differ meaningfully between users?

We compared the deferral scores of all users together, and of pairs of individual users to determine whether the distribution of deferral scores caused by the initial query is dependent on the user. Based on our previous results, we used a 45 task burn-in. We found using a Kruskal-Wallis test that there was a significant effect of user on deferral scores ($p < 0.05$) and a Mann-Whitney U test showed that the distributions were significantly different ($p < 0.05$) for 79 of the 300 user pairings.

As we see in Figure 7, the distributions of scores may be different even if the final achieved error is the same: the solid pink line has many more high-certainty examples balanced by more low-certainty examples, while the dashed pink line is more evenly distributed. A Mann-Whitney U test revealed that both pairs are significantly different ($p < 0.05$). This finding shows that, whether or not we control for error, users can produce significantly different deferral scores. When this is considered together with the formulation described in Section 3.2 and the relationship between error, deferral rate, and satisfaction shown in RQ1, it shows that *deferral criteria must be set based on the individual who is using the AI agent.*

## RQ4: How do users respond when an inference has been deferred?

*Analysis of Model Response.* We begin by analyzing the interaction of the user and deep-learned model after deferral in terms of deferral score and error. Since the data are paired and deferral only begins after the $30^{th}$ task, we evaluated all deferral responses without regard to the burn-in. Although the standard approach

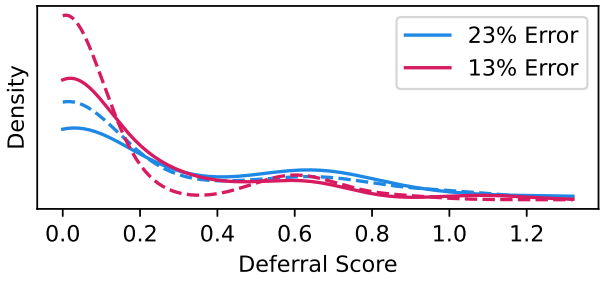


**Figure 7: Kernel Density Estimate plots of deferral score frequency for four different users. Despite users with corresponding colors having the same overall error, both pairs are visibly and statistically ($p < 0.05$ by Mann-Whitney U) different. Kernel Bandwidth set by Scott's rule.**

| | Example | Count |
|---|---|---|
| Identical | bottom left bed<br>→ bottom left bed | 7 |
| Rephrase | giraffe on the left<br>→ left giraffe | 123 |
| Same Detail | donut with chocolate sprinkles<br>→ donut at the bottom | 169 |
| More Detail | the car on the right<br>→ the car covered by snow on the right hand side | 66 |
| Less Detail | plants in green basket behind roses<br>→ plants in green basket | 91 |

**Table 1: Types of deferral responses and quantity of each seen in our experiment.**

of reformulation used in conversational virtual assistants [33, 71] makes the implicit assumption that the human would provide a better utterance after deferral, we found that the deferral response was of lower quality (from the model's perspective) than the initial query: the mean output entropy in aggregate increased from 0.204 to 0.239 with significance (Wilcoxon Signed-Rank test, $p < 0.05$) and two users showed a statistically significant difference in output entropy between the initial query and deferral response, both of whom had a higher entropy after deferral (Wilcoxon Signed-Rank test, $p < 0.05$). Although we could not show significance, the error in aggregate for the first input (19.82%) was lower than the error for the second input (23.39%).

Although the deferral response was of lower quality than the initial input, we found that by using an aggregation function we could still reduce error over the deferral free condition: error decreased from 19.82% to 17.37% after deferral, 30 out of 89 (33.71%) incorrect answers were corrected, and 19 out of 360 (5.28%) correct answers were made incorrect. The results were similar when data from the burn-in was included: error decreased from 19.45% to 18.01%, 25 out of 72 (34.72%) incorrect inferences were made correct, while 18 out of 289 (6.23%) of correct inferences were made incorrect. Although this reduction in overall error suggests the importance of a well-chosen aggregation function, McNemar's test did not reveal significance ($p > 0.05$).

This finding provides two important insights into these kinds of problems. First, since the deferral response is generally of lower quality than the initial query, the naive reformulation approach [71] is insufficient: not only will error increase with an increased deferral rate after reaching a minimum [48], but deferral rates greater than this minimum may actually have a higher error than the deferral free condition. Therefore, it is critical to maintain state and use a meaningful aggregation function. Second, $p(s_2|s_1, u)$ can not be approximated using $p(s1|u)$, meaning that prior to being able to target an error via deferral using Equation 5, we must perform multiple deferrals to characterize $p(s_2|s_1, u)$.

*Input-Space Analysis.* In addition to examining how the model responds to initial queries and deferral responses, it is informative to characterize how users respond in the input space. On an individual basis, there were three users with a statistically significant difference in sentence length (Wilcoxon Signed-Rank test, $p < 0.05$), none of whom also had a statistically significant difference in deferral score. All of these users had a greater length for their deferral response. To provide further understanding of deferral responses, we grouped all 449 examples[2] into five broad categories: *Identical*, where the first phrase was re-used without change; *Rephrase*, where the semantic meaning and detail remained unchanged despite a change in wording; *Same Detail*, where there were meaningful semantic differences but roughly the same amount of overall information; *More Detail*, where the second input either added data to the previous phrase or used a clearly more detailed independent phrase; and *Less Detail*, where the deferral response contained less information than the initial query.

We show the number of times each category occurred in Table 1: most of the deferral responses were of equivalent detail, with users slightly preferring to modify semantics (same detail) over syntax (rephrase). This large proportion of rephrasing events (25.8% of deferral responses) suggests that methods used for extracting deferral responses from datasets, such as random sampling [47] or minimum word overlap [34], are likely insufficient for many settings. Although no participant systematically produced shorter responses to a statistically significant degree, aggregate analysis suggests that users believe it to be more likely that the model will understand less information better than more, consistent with previous findings that humans use shorter messages with chatbots [36]. Interestingly, we did not find any cases where the deferral response was ambiguous without the initial referring expression (*e.g.,* the leftmost flowers → the yellow ones), meaning the increased entropy after deferral was likely to be due to the aforementioned shortening of messages or the fact that training data [39] consisted entirely of initial requests. Additionally, since we did not explicitly state that the AI remembered the previous interaction, this suggests that users assume the AI agent does not remember previous queries, and any deferral method with memory should therefore make this feature explicit to the user.

## RQ5: Does knowing the user provide additional information about the mapping between probability of error and deferral score?

From the formulation described in Equation 5, we see that it is important to consider the relationship between deferral score and the probability of error—$p(e|s_1, u)$ and $p(e|s_2, u)$—if we want to target an error. Due to the large number of samples required to build this distribution, it would meaningfully reduce the calibration time required to set deferral criteria if these distributions were independent of the user (*i.e.,* $p(e|s_1) = p(e|s_1, u) \ \forall \ u$). To determine if this is the case, we measured whether the deferral score gives us more information about the probability of error if it is conditioned on a user. We measured this using Mutual Information (MI), which describes the dependency of two random variables and has been used for tasks such as measuring the quality of fused images [32] and choosing network weights to prune [25], and compared the strength of this relationship when it is and isn't conditioned on an individual user using a permutation test:

[2]Due to a connectivity error, one user had one fewer deferral, leading to 449 responses instead of 450.

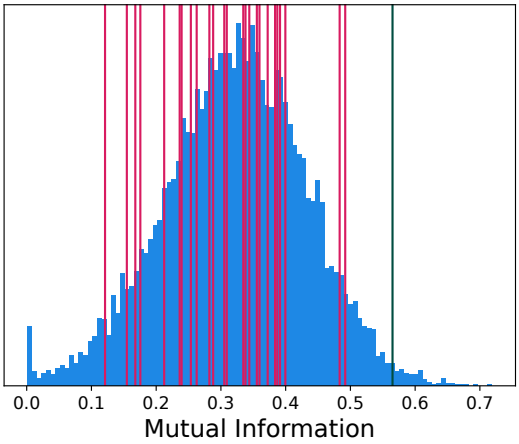


**Figure 8: The mutual information conditioned on individual users (pink lines) superimposed on a distribution of unconditioned mutual information (blue histogram). The green line represents the p < 0.05 significance threshold.**

(1) Draw 10,000 random sets of 75 score-error pairs (120 tasks less the 45 task burn-in) and calculate the mutual information of each one using the EDGE estimator [61].
(2) For every individual user, calculate the mutual information between the deferral score and the error.
(3) Compare every individual user's MI to the distribution of randomly generated MIs.

We see the results of this in Figure 8. We see that none of our 25 evaluated users allowed us to reject the null hypothesis that knowing the user does not increase mutual information ($p > 0.05$). Though this phenomenon would benefit from further study, this finding suggests that $p(e|s_n)$ is independent of the user and *we can use dataset-based model calibration to estimate the probability of error given a deferral score.* Doing this would dramatically decrease the time required to set deferral criteria over characterizing the model based on individual user interactions.

## 6 SETTING DEFERRAL CRITERIA

Our work has thus far shown that user satisfaction is dependent on both deferral rate and error and that users are unique, but has not explicitly shown that it is better to set deferral criteria based on the data from individual participants. To investigate this we set deferral rate and threshold based on two objectives:

(1) Set deferral criteria to bring the deferral rate closest to the target value. (*Minimizing Absolute Error*)
(2) Set deferral criteria to produce an upper bound on the deferral rate. (*Upper Bounding*)

We considered three calibration datasets, all of which are evaluated on the final 38 human inputs (half of the tasks remaining for an individual after burn-in):

- RefCOCO: set deferral criteria using phrases from the RefCOCO dataset [39]. This dataset was constructed using a crowdsourced two-player human-to-human game, where the benefit of increased size may be outweighed by the difference in human-to-human communication [36]. We used only images that met the criteria defined in our experimental setup, and removed all images that were seen by the user for whom we are setting deferral criteria.
- Multi-User: set deferral criteria using phrases collected from other users in the experiment. This dataset is slightly smaller than RefCOCO, but collected in the same setting as the test data. We remove from the calibration set all phrases from prior to the burn-in (the first 45), as well as all images that were seen by the target user.
- Individual: set the deferral criteria based on the first half (37) of phrases after the burn-in. Although the calibration set is much smaller, it will also capture the user's behavior much more accurately.

*Minimizing Absolute Error.* The method for minimizing the absolute error with respect to a deferral rate is to find the value of the appropriate percentile in the calibration set. We see the result of this in Table 2: no method has a mean of greater than one standard error from another. Although the performance of RefCOCO and Multi-User deferral criteria is likely stable—having approximately 2,900 and 1,650 samples, respectively—deferral criteria based on an individual may improve with a longer calibration period. In other words, while setting deferral criteria using an individual does not improve over aggregate datasets in this analysis, a longer interaction period may change this finding.

*Upper Bounding.* Given our finding that user satisfaction is linked to both deferral and error rates, it makes sense to set an upper bound on the the respective value (*i.e.,* I want the user to be at least this happy) instead of simply attempting to match the desired value as closely as possible at the risk of great dissatisfaction for some users. This is particularly critical for faster calibrations, as minimizing the absolute error does not consider the number of examples in the calibration set.

To upper bound the error rate, we use the finding of Gascuel & Caraux [26] that when $\bar{p}$ verifies:

$$\delta = \sum_{i=0}^{k} \binom{n}{i} \bar{p}^i (1-\bar{p})^{n-i}, \tag{9}$$

then

$$p(p - \bar{p} \geq 0) \leq \delta, \tag{10}$$

where $p$ is the true probability of deferral for the proposed criteria, $n$ is the number of examples, $k$ is the number of deferrals for a given deferral criteria, and $\delta$ is our desired confidence. Like Geifman & El-Yaniv [27] do for the selective classification task, we solve this using a binary search across deferral criteria (the threshold, $t$) with $\delta = 0.05$.

For this goal, deferral criteria produced by only examining the individual unambiguously performs better (Table 3). For RefCOCO and Multi-User, the deferral criteria is set with high confidence due to the size of the calibration set, but is incorrect due to the differences in score distributions between individuals. In other words, thresholding based on an individual's score distribution is necessary for producing accurate upper bounds, regardless of the calibration set size.

| | 0.1 | 0.2 | 0.3 |
|---|---|---|---|
| RefCOCO | 3.82 ± 0.63 | 5.66 ± 0.88 | 8.41 ± 0.90 |
| Multi-User | 4.22 ± 0.67 | 5.94 ± 0.88 | 6.97 ± 1.14 |
| Individual | 3.97 ± 0.70 | 6.95 ± 0.91 | 6.79 ± 1.09 |

**Table 2: Mean absolute error when targeting deferral rates of 0.1, 0.2, and 0.3. Displayed tolerances are standard error.**

| | 0.1 | 0.2 | 0.3 |
|---|---|---|---|
| RefCOCO | 7 | 9 | 7 |
| Multi-User | 10 | 10 | 9 |
| Individual | 0 | 1 | 2 |

**Table 3: Number of violations when upper bounding deferral rates of 0.1, 0.2, and 0.3.**

## 7 DISCUSSION AND FUTURE WORK

In this work, we provided a human-centered view of deferred inference with deep networks. Through a study of 25 users, we examined not only whether error is reduced by the addition of a deferral mechanism—as in previous work—but also the nuances of the interaction between individual users and deep learning models. Guided by the formulation of Section 3.2, we report several important findings. Most broadly, we find that 1) satisfaction is dependent on both error and deferral rates (RQ1), and 2) setting deferral criteria to target an error or deferral rate must consider qualities of an individual (RQ3). The second finding is reinforced in practice by an evaluation of different methods for setting deferral criteria: despite having two orders of magnitude less data, deferral criteria set with user-specific data perform the same or better than those set on large datasets containing many individuals. We additionally find that it is critical to characterize the deferral response separately from the initial query (RQ4) but that we can characterize the model's calibration—the relationship between score and error—independently from the individual (RQ5).

Though deep neural networks are inherently unpredictable, we believe that the findings of our work are sufficiently general to extend to other relevant applications. Many are likely to be model-agnostic: people have subtly different linguistic preferences, and the ways in which they change their language after deferral is a function of the human's perception of the model, not the model itself. The broad concepts for setting deferral criteria as a threshold on a deferral score is also likely to generalize, though the deferral function itself will have to change if the output format is different: visual question answering [1] often uses a softmax output [11, 18, 63], but there is no trivial equivalent to entropy in, for example, the bounding box output of a visual object tracker [42].

In addition to other applications, future work should consider a study with longer interactions for each user, potentially across multiple sessions in a real-world scenario. Such a study will allow us to answer three other important questions: 1) can deferral with custom criteria reduce error to a statistically significant degree? Our choice to use random deferral instead of basing our deferral criteria on entropy led to many high-certainty (and correct) answers being deferred. Because of this, we could not show statistically significant improvement even though it was much more likely that the post-deferral answer was correct given an incorrect pre-deferral answer (33.71%) than the opposite (5.28%). 2) Can we target an error? Our evidence suggests that we can use datasets to estimate the probability of error given a deferral score—mitigating some concerns about interaction length—but the nature of Bernoulli variables makes it difficult to produce meaningful evaluations at small sample sizes. 3) Is there a longer-term shift in user behavior that was not captured in our study? In other words, we may need to re-calibrate the deferral criteria over time to maintain our target value. Future work should also solidify the exploration of user preference: how long does the the plateau in Figure 4-B last? Can we develop a Pareto front that directly models the trade-off between error and deferral rate?

The ethical implications of this work roughly track those of deep learning in general—if the model has meaningful biases [35, 55, 85], those biases will still be reflected in its output, and it is not recommended to use such methods for sensitive applications. However, the straightforward implementation of deferred inference allows three meaningful opportunities for improvement in this space that merit further study. First, by respecting the standard single-input-single-output formulation of deep neural networks, the method described in this work allows us to rapidly deploy new architectures that may have a greater ability to compensate for such biases [35], without needing to develop additional learned models to enable multiple human inputs. Second—though we recommend additional investigation—inferences that make incorrect assumptions are likely to have higher deferral scores, or deferral scores can be produced specifically for this purpose, potentially allowing the user to correct such problems before they have an effect. Similarly, user-specific characteristics, such as dialects, that are not well understood by the model are likely to be deferred, allowing the user to resolve the issue. Although this comes at the cost of increased user effort, this is likely preferable to the practices of confidently returning an incorrect answer or returning no answer at all.

## 8 CONCLUSION

Deferred inference is an intuitive and effective way to improve performance of a pre-existing model, but deferral criteria for such methods are typically set only on the model's confidence. In doing so, they ignore both user satisfaction and qualities of the user. In this work, we demonstrated the importance of considering these user-dependent characteristics in deferred inference: satisfaction is tied to both error and deferral rate, and both of these values are dependent on the individual. Through these findings, we lay necessary groundwork for a simple method of interaction with deep networks that can be rapidly implemented to improve performance and user satisfaction.

## ACKNOWLEDGMENTS

Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

# REFERENCES


[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*. IEEE Press, Santiago, Chile, 2425–2433.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the 2019 AAAI Conference on Human Computation and Crowdsourcing*. AAAI Press, Orlando, Florida, USA, 2–11.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *Proceedings of the 2019 AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii, USA, 2429–2437.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM Press, Yokohama, Japan, 1–16.

[5] Giuseppe Bevacqua, Jonathan Cacace, Alberto Finzi, and Vincenzo Lippiello. 2015. Mixed-Initiative Planning and Execution for Multiple Drones in Search and Rescue Missions. In *Proceedings of the 2015 International Conference on Automated Planning and Scheduling*. AAAI Press, Jerusalem, Israel, 315–323.

[6] Nilava Bhattacharya, Qing Li, and Danna Gurari. 2019. Why Does a Visual Question Have Different Answers?. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*. IEEE Press, Seoul, South Korea, 4270–4279.

[7] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 2010 Annual ACM Symposium on User Interface Software and Technology*. ACM Press, New York, New York, USA, 333–342.

[8] Jeffrey P. Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. 2010. VizWiz::LocateIt - enabling blind people to locate objects in their environment. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE Press, San Francisco, California, USA, 65–72.

[9] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. 2022. Role of Human-AI Interaction in Selective Prediction. In *Proceedings of the 2022 AAAI Conference on Artificial Intelligence*. AAAI Press, Virtual, 5286–5294.

[10] J. Cacace, A. Finzi, V. Lippiello, M. Furci, N. Mimmo, and L. Marconi. 2016. A control architecture for multiple drones operated via multimodal interaction in search & rescue mission. In *Proceedings of the 2016 IEEE International Symposium on Safety, Security, and Rescue Robotics*. IEEE Press, Lausanne, Switzerland, 233–239.

[11] Remi Cadene and Corentin Dancette. 2019. RUBi: Reducing Unimodal Biases for Visual Question Answering. In *Proceedings of the 2019 Conference on Advances in Neural Information Processing Systems*. Curran Associates, Vancouver, British Columbia, Canada, 839–850.

[12] 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 Press, Glasgow, Scotland, UK, 14.

[13] Felix Carros, Johanna Meurer, Diana Löffler, David Unbehaun, Sarah Matthies, Inga Koch, Rainer Wieching, Dave Randall, Marc Hassenzahl, and Volker Wulf. 2020. Exploring Human-Robot Interaction with the Elderly: Results from a Ten-Week Case Study in a Care Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM Press, Honolulu, HawaiI, USA, 1–12.

[14] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 2015 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, Sydney, New South Wales, Australia, 1721–1730.

[15] Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. In *Proceedings of the 2006 International Conference on Intelligent Tutoring Systems*. Springer, Jhongli, Taiwan, 164–175.

[16] Minsuk Chang, Mina Huh, and Juho Kim. 2021. RubySlippers: Supporting Content-based Voice Navigation for How-to Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM Press, Yokohama, Japan, 97:1–97:14.

[17] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to Design Voice Based Navigation for How-To Videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM Press, Glasgow, Scotland, UK, 701–712.

[18] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Proceedings of the 2020 European Conference on Computer Vision*. Springer, Virtual, 104–120.

[19] C. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16, 1 (Jan. 1970), 41–46.

[20] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. 2016. Boosting with Abstention. In *Proceedings of the 2016 Conference on Advances in Neural Information Processing Systems*. Curran Associates, Barcelona, Spain, 1660–1668.

[21] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can I help you with?": infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 2017 International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM Press, Vienna Austria, 43:1–43:12.

[22] Martin Danelljan, Luc Van Gool, and Radu Timofte. 2020. Probabilistic Regression for Visual Tracking. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE PRess, Virtual, 7183–7192.

[23] Giorgio Fumera and Fabio Roli. 2002. Support Vector Machines with Embedded Reject Option. In *Proceedings of the 2002 Pattern Recognition with Support Vector Machines Workshop*. Springer Berlin Heidelberg, Niagara Falls, Ontario, Canada, 68–82.

[24] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 2016 International Conference on Machine Learning*. PMLR, New York, New York, USA, 1050–1059.

[25] Madan Ravi Ganesh, Jason J. Corso, and Salimeh Yasaei Sekeh. 2021. MINT: Deep Network Compression via Mutual Information-based Neuron Trimming. In *Proceedings of the 2020 International Conference on Pattern Recognition*. Springer, Virtual, 8251–8258.

[26] Olivier Gascuel and Gilles Caraux. 1992. Distribution-free performance bounds with the resubstitution error estimate. *Pattern Recognition Letters* 13, 11 (Nov. 1992), 757–764.

[27] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Proceedings of the 2017 Conference on Advances in Neural Information Processing Systems*. Curran Associates, Long Beach, California, USA, 4878–4887.

[28] Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 2019 International Conference on Machine Learning*. ACM Press, Long Beach, California, USA, 2151–2159.

[29] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 2017 International Conference on Machine Learning*. PMLR, Sydney, New South Wales, Australia, 1321–1330.

[30] 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*. ACM Press, Denver, Colorado, USA, 3511–3522.

[31] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Press, Salt Lake City, Utah, USA, 3608–3617.

[32] Mohammad Haghighat and Masoud Amirkabiri Razian. 2014. Fast-FMI: Non-reference image fusion metric. In *Proceedings of the 2014 IEEE International Conference on Application of Information and Communication Technologies*. IEEE Press, Paris, France, 1–3.

[33] Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and Predicting Voice Query Reformulation. In *Proceedings of the 2015 ACM International on Conference on Information and Knowledge Management*. ACM Press, Melbourne, Victoria, Australia, 543–552.

[34] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. 2018. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation*. IEEE Press, Brisbane, Queensland, Australia, 3774–3781.

[35] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of the 2018 European Conference on Computer VIsion*. Springer International Publishing, Munich, Germany, 793–811.

[36] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49 (Aug. 2015), 245–250.

[37] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM Press, Montreal, Quebec, Canada, 1–6.

[38] Suyog Dutt Jain and Kristen Grauman. 2016. Click Carving: Segmenting Objects in Video with Point Clicks. In *Proceedings of the 2016 AAAI Conference on Human Computation and Crowdsourcing*. AAAI Press, Austin, Texas, USA, 89–98.

[39] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Doha, Qatar, 787–798.

[40] Jong-Wook Kim, Young-Lim Choi, Sang-Hyun Jeong, and Jeonghye Han. 2022. A Care Robot with Ethical Sensing System for Older Adults at Home. *Sensors* 22, 19 (Oct. 2022), 7515.

[41] Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine* 4, 1 (Dec. 2021), 4.

[42] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernández, Georg Nebehay, Fatih Porikli, and Luka Čehovin. 2016. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 11 (Nov. 2016), 2137–2155.

[43] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM Press, New Orleans, Louisiana, USA, 54:1–54:18.

[44] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM Press, New Orleans, Louisiana, USA, 462:1–462:15.

[45] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. 2017. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports* 7, 1 (Dec. 2017), 1–14.

[46] Stephan J Lemmer and Jason J Corso. 2021. Ground-Truth or DAER: Selective Re-Query of Secondary Information. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*. IEEE Press, Virtual, 703–714.

[47] Stephan J. Lemmer and Jason J. Corso. 2023. Evaluating and Improving Interactions with Hazy Oracles. In *Proceedings of the 2023 AAAI Conference on Artificial Intelligence*. AAAI Press, Washington, District of Columbia, USA, 9.

[48] Stephan J. Lemmer, Jean Y. Song, and Jason J. Corso. 2021. Crowdsourcing More Effective Initializations for Single-Target Trackers Through Automatic Re-querying. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM Press, Virtual, 391:1–391:13.

[49] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions. In *Proceedings of the 2018 European Conference on Computer Vision*. Springer, Munich, Germany, 570–586.

[50] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 2014 European Conference on Computer Vision*. Springer, Zurich, Switzerland, 740–755.

[51] Brian Lucena. 2018. Spline-Based Probability Calibration.

[52] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM Press, San Jose, California, USA, 5286–5297.

[53] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM Press, Denver, Colorado, USA, 5988–5999.

[54] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. 2017. The Promise of Premise: Harnessing Question Premises in Visual Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 926–935.

[55] Varun Manjunatha, Nirat Saini, and Larry S. Davis. 2019. Explicit Bias Discovery in Visual Question Answering Models. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE PRess, Long Beach, California, USA, 9554–9563.

[56] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Press, Las Vegas, Nevada, USA, 11–20.

[57] Oier Mees and Wolfram Burgard. 2020. Composing Pick-and-Place Tasks By Grounding Language. In *Proceedings of the 2020 International Symposium on Experimental Robotics*. Springer, La Valletta, Malta, 491–501.

[58] Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. 2018. Attended Temperature Scaling: A Practical Approach for Calibrating Deep Neural Networks.

[59] Caio Mucchiani, Pamela Cacchione, Michelle Johnson, Ross Mead, and Mark Yim. 2021. Deployment of a Socially Assistive Robot for Assessment of COVID-19 Symptoms and Exposure at an Elder Care Setting. In *Proceedings of the 2021 IEEE International Conference on Robot & Human Interactive Communication*. IEEE Press, Virtual, 1189–1195.

[60] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 2018 Annual ACM Symposium on User Interface Software and Technology*. ACM Press, Berlin Germany, 189–199.

[61] Morteza Noshad, Yu Zeng, and Alfred O. Hero III. 2019. Scalable Mutual Information Estimation using Dependence Graphs. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Press, Brighton, United Kingdom, 2962–2966.

[62] Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. 2018. Grounding Robot Plans from Natural Language Instructions with Incomplete World Knowledge. In *Proceedings of the 2018 Conference on Robot Learning*. PMLR, Zurich, Switzerland, 714–723.

[63] Amelia Elizabeth Pollard and Jonathan L. Shapiro. 2020. Visual Question Answering as a Multi-Task Problem.

[64] Prakruthi Prabhakar, Nitish Kulkarni, and Linghao Zhang. 2018. Question Relevance in Visual Question Answering.

[65] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. 2019. Direct Uncertainty Prediction for Medical Second Opinions. In *Proceedings of the 2019 International Conference on Machine Learning*. ACM Press, Long Beach, California, USA, 5281–5290.

[66] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, USA, 2383–2392.

[67] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, San Francisco, California, USA, 1135–1144.

[68] Kelly Rivers, Erik Harpstead, and Ken Koedinger. 2016. Learning Curve Analysis for Programming: Which Concepts do Students Struggle With?. In *Proceedings of the 2016 ACM Conference on International Computing Education Research*. ACM Press, Melbourne, Victoria, Australia, 143–151.

[69] Lucas Rosenblatt, Patrick Carrington, Kotaro Hara, and Jeffrey P. Bigham. 2018. Vocal Programming for People with Upper-Body Motor Impairments. In *Proceedings of the 2018 International Web for All Conference*. ACM Press, Lyon, France, 30:1–30:10.

[70] Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. 2018. The Elephant in the Room.

[71] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2017. Predicting Causes of Reformulation in Intelligent Assistants. In *Proceedings of the 2017 Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken, Germany, 299–309.

[72] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* 128, 2 (Feb. 2020), 336–359.

[73] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. 2022. Correcting Robot Plans with Natural Language Feedback. In *Proceedings of the 2022 Conference on Robotics: Science and Systems*. MIT Press, New York, New York, USA, 1–12.

[74] Mohit Shridhar and David Hsu. 2018. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In *Proceedings of Robotics: Science and Systems 2018*. MIT Press, Pittsburgh, Pennsylvania, United States, 1–9.

[75] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the 2014 International Conference on Learning Representations*. OpenReview, Banff, Alberta, Canada, 10.

[76] Ryan Szeto and Jason J. Corso. 2017. Click Here: Human-Localized Keypoints as Guidance for Viewpoint Estimation. In *Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision*. IEEE Press, Venice, Italy, 1604–1613.

[77] Jorge Sánchez, Mauricio Mazuecos, Hernán Maina, and Luciana Benotti. 2022. What kinds of errors do reference resolution models make and what can we learn from them?. In *2022 Findings of the Association for Computational Linguistics*. ACL Press, Seattle, Washington, USA, 1971–1986.

[78] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-Dialog Navigation. In *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Virtual, 394–406.

[79] Kohei Uehara, Nan Duan, and Tatsuya Harada. 2022. Learning To Ask Informative Sub-Questions for Visual Question Answering. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE Press, New Orleans, Louisiana, USA, 4681–4690.

[80] Jasper R. R. Uijlings, Mykhaylo Andriluka, and Vittorio Ferrari. 2020. Panoptic Image Annotation with a Collaborative Assistant. In *Proceedings of the 2020 ACM International Conference on Multimedia*. ACM Press, Virtual, 3302–3310.

[81] K. R. Varshney. 2011. A risk bound for ensemble classification with a reject option. In *2011 IEEE Statistical Signal Processing Workshop*. IEEE Press, Nice, France, 769–772.

[82] David Widmann, Fredrik Lindsten, and Dave Zachariah. 2019. Calibration tests in multi-class classification: A unifying framework. In *Proceedings of the 2019 Conference on Advances in Neural Information Processing Systems*. Curran Associates, Vancouver, British Columbia, Canada, 12236–12246.

[83] Xuhai Xu, Jun Gong, Carolina Brum, Lilian Liang, Bongsoo Suh, Kumar Gupta, Yash Agarwal, Laurence Lindsey, Runchang Kang, Behrooz Shahsavari, Tu Nguyen, Heriberto Nieto, Scott E. Hudson, Charlie Maalouf, Seyed Mousavi, and Gierad Laput. 2022. Enabling hand gesture customization on wrist-worn devices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM Press, New Orleans, Lousiana, USA, 496:1–496:19.

[84] Takashi Yamamoto, Koji Terada, Akiyoshi Ochiai, Fuminori Saito, Yoshiaki Asahara, and Kazuto Murase. 2019. Development of Human Support Robot as the research platform of a domestic mobile manipulator. *ROBOMECH Journal* 6, 1 (Dec. 2019), 4.

[85] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM Press, Barcelona Spain, 547–558.

[86] Jennifer Zamora. 2017. I'm Sorry, Dave, I'm Afraid I Can't Do That: Chatbot Perception and Expectations. In *Proceedings of the 2017 International Conference on Human Agent Interaction*. ACM Press, Bielefeld, Germany, 253–260.

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

[88] Yaxi Zhao, Razan Jaber, Donald McMillan, and Cosmin Munteanu. 2022. "Rewind to the Jiggling Meat Part": Understanding Voice Control of Instructional Videos in Everyday Tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM Press, New Orleans LA USA, 58:1–58:11.