Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models

Reviewed on OpenReview:

Abstract

Human experiences are complex and subjective. This subjectivity is reflected in the way people label images for machine vision models. While annotation tasks are often assumed to deliver objective results, this assumption does not allow for the subjectivity of human experience. This paper examines the implications of subjective human judgments in the behavioral task of labeling images used to train machine vision models. We identify three primary sources of ambiguity: (1) depictions of labels in the images can be simply ambiguous, (2) raters' backgrounds and experiences can influence their judgments and (3) the way the labeling task is defined can also influence raters' judgments. By taking steps to address these sources of ambiguity, we can create more robust and reliable machine vision models.

Keywords: Disagreements, Ambiguity, Machine vision

1 Introduction

Computer vision models rely on human annotations, and the default assumption when creating training and evaluation datasets is often that there is a single correct answer about what concepts or objects are present in an image. Though there is growing acceptance that human disagreements are common with respect to inherently ambiguous data Kairam and Heer (2016), the role of human disagreements as a general property of *any* annotation task is much less accepted. In image annotation, even the annotation of concrete concepts (e.g., *bird*) in clear, high quality, unobscured imagery can lead to disagreement between raters that we should seek to understand. The interplay of annotator, concept, and image characteristics in labeling tasks should inform how we analyze human ratings, leverage disagreement insights to train and evaluate models, and translate findings into best practices.

To understand individual human behavior in image annotation, we focus on large label space models for computer vision. *Large label space models* are machine vision models that predict the probabilities of many entities in an image, in contrast to *binary classification models* that predict the presence or the absence of a single entity and *segmentation models* that identify pixels corresponding to an entity. Most image models require labeled training data to learn to classify accurately (e.g., Ji et al. (2019)). This requirement typically consists of a training set of images labeled with their contents, usually by human annotators. For example, to learn to classify birds in images, a large label space model would need to see many (usually at least tens of thousands) of images of birds, depicted in a range of different environments and positions with the inclusion of rare species. Human annotators are employed to label each image, providing the ”ground truth” needed to train the model.

We know, however, that *human raters disagree.* Bird experts may disagree on which species of bird an image belongs to. Non-experts may be unsure about taxonomic classifications of certain bird species. People can disagree whether the concept of "bird" applies in a given case (e.g., *pictures* of birds). Some reasons, like poor image quality, can indicate problems with a specific image. However, many cases of human disagreements are due to ambiguity in the label or the labeling task. Label ambiguity can arise from many factors, including similar-looking labels (*birds* and *bats* look similar), regional naming differences (*robin* in the US, vs. *redbreast* in the UK), and different understandings of the task. Label ambiguity is a challenge for machine vision models because it can lead to inaccurate predictions. For example, if a machine vision model is trained on a dataset of *bird* images with ambiguous labels, it may not be able to accurately identify birds in new images (see Karimi et al. (2020) for an analysis of the impact of label noise on medical image analysis models).

In order to better understand the human factors that influence label ambiguity on large label space model performance, we developed an open data challenge to crowdsource *adversarial image-labels pairs* for machine vision models. In this online challenge, participants competed to identify edge case images that state-of-the-art machine vision models might incorrectly classify. The goal was to understand systematic failures of these models, with an eye towards augmenting the data used to train these models to better cover such failure cases.Image-label pairs collected during the challenge were tested against multiple state-of-the-art classification models and surfaced (i) pairs with clear human-machine disagreements and (ii) pairs where multiple *human* annotators couldn't reach clear agreement. This challenge required a data analysis strategy designed to identify patterns in the misclassified images to better understand the human factors that contribute to label ambiguity and to develop new mitigation methods. The adversarial data from this challenge and the resulting analysis have the potential to make a significant contribution to the development of more robust and reliable large label space models. In this paper, we present the results of the public adversarial data challenge, analyze the ambiguities in the resulting data, and organize them into a theoretical framework to provide recommendations for human annotation and data collection policies that best address the types of ambiguities we observed.

2 Adversarial data challenge

The adversarial data challenge ran online for four months, under the CrowdCamp umbrella of the HCOMP 2021. The challenge used the Open Image Dataset[1] (OID V4; Krasin, 2017) as source material. It contains ∼9M images annotated with 20k possible image-level labels, object bounding boxes and segmentation masks. Importantly, the labels, bounding boxes, and segmentation masks are provided by a machine, with only a small portion verified by humans. The challenge was designed on the premise that, likely, the machine labeler makes mistakes, these mistakes are systematic, and studying systematic machine failures can improve machine labelers in the future. In this challenge, we aimed to identify adversarial image-label pairs in OID V4 that would yield human-model disagreement.

Challenge participants examined the machine-labeled subset of OID V4 images, focusing on a selected set of 23 entities - *Bird, Canoe, Lipstick, Chopsticks, Muffin, Pizza, Croissant, Child, Smile, Selfie, American football, Athlete, Physician, Nurse, Teacher, Chef,*

[1] `https://storage.googleapis.com/openimages/web/factsfigures_v4.html`

Firefighter, Coach, Construction Worker, Bus driver, Funeral, Thanksgiving, or Graduation - and submit image-label pairs where they thought the image classification machine algorithm was wrong. Limiting the label set to 23 was necessary to make the scope of the competition tractable—human participants were unlikely to be able to examine all 20k labels in the OID. These 23 labels were selected to represent a neutral (non-controversial, non-sensitive) set of topics across different types: 8 objects, 3 events, 9 roles and professions, and 3 abstract concepts. Another criteria for selection was to have a good representation of different levels of ambiguity of the label, e.g. "child" is a broad concept and could be interpreted in different ways; "athlete" could mean different things for different cultures; "physician" and "nurse" could have ambiguous visual representations.

Ten individuals submitted image-label pairs to the challenge, submitting more than 14,000 image-label pairs. Of these, 13,683 image-label pairs were "valid" (i.e., the pairs were drawn from OID V4 and used one of the 23 challenge labels). After removing duplicate submissions, 10,668 unique pairs remained. Participants could choose for which labels of the 23 to submit and how many images. The 10,668 unique image-label pairs were further validated by engaging two globally-diverse crowds of human annotators in three different locales and two in-house experts (described the Methods section). The image-label pairs were also submitted to six machine vision models to examine how human judgements aligned with state of the art model judgements and to identify cases of human-model disagreements. The challenge data is on github (see reproducibility statement for non-anonymous link); additional human annotations collected for this study will be made available after review.

3 Methods

Here, we provide a detailed description of the materials (datasets and models) used, annotation task procedures, task annotators, and data analysis and score computation decisions that we made in arriving at the results, all summarized in Appendix Figure 1.

3.1 Materials

The challenge dataset. As described above, the challenge dataset was composed of 10,668 unique submissions made by challenge participants. Appendix Figure 4 shows the distribution of images across all 23 target labels - most images were submitted for the label 'bird' (26% of the data) with an exponential long-tail distribution across all other labels (e.g., seven labels with between 500-1100 images per label, nine with between 140-350 images per label and six labels with 100 or fewer images per label).

Vision models. To provide machine labels of each challenge dataset image, we used an ensemble of six machine vision models, each of which were state-of-the-art when they were released. These models are all non-public variants of the InceptionV2-based image classifier Ioffe and Szegedy (2015) developed in the period of 2015-2022 (including models used in OID-V4 and OID-V3 Krasin (2017), publicly available through Open Images Dataset).

Model error dataset. Based on human annotation Task 1 and qualitative validation by experts, we constructed a subset of 8,326 image-label pairs to have labeled by humans in Task 2. Image-label pairs included in Task 2 met at least one of the following criteria: (i) at least one vision model disagreed with the human majority vote from Task 1, or (ii) there was significant disagreement among the human annotators in Task 1. This smaller dataset allows for a targeted qualitative analysis of the reasons for human-model disagreements.

3.2 Annotation task procedure

Human annotation task 1—Label verification. In Task 1, annotators indicated whether a given label applied to an image for each image-label pair in the challenge dataset. No specific training was provided to annotators before beginning the task, as the task was injected into a general purpose image-label validation system used by a professional rater pool to perform a variety of tasks other than this one. For each example, 19 annotators viewed a single image and selected one of three answer options indicating whether a given label applies to that image, does not apply, or they are unsure (Appendix Figure 2).

Human annotation task 2—Model error verification. In Task 2, annotators examined the model error dataset (8,326 image-label pairs from the challenge dataset with human and machine labelers disagreement from Task 1). For each example, 14 annotators saw a machine label produced for an image, and they indicated whether the model was correct or not (Appendix Figure 3.A). Guidelines (presented to annotators before starting) included definitions of seven categories of model error that were identified by experts in a qualitative analysis of a subset of the model error dataset (see Data Analysis section). Annotators answered two questions about each item: (i) whether the machine prediction indicated correctly whether the label was present in the image or not (Figure 3.B), and (ii) in the case of model error, select one out of seven possible error types (Figure 3.C). Annotators were not given any information about how the "machine prediction" was constructed in order to avoid biasing them towards agreeing or disagreeing with the machine prediction.

3.3 Annotators

Data submitted by challenge participants was validated three times—twice by paid annotators and once by members of the research team. The paid annotators were recruited from professional rater pools and had prior experience in data annotation tasks. To ensure that the annotations on the image-label pairs reflected a range of human perspectives, particularly because we expected that the examples would be especially challenging, we recruited raters from different geographic locales (US, Canada, and India). We selected these locales because they have English as a dominant language and are common locales for recruiting annotators. Details about raters in each task and their selection is outlined in Appendix D.

3.4 Scoring

Merging task 1 and task 2 human labels. Tasks 1 and 2 both asked annotators if a label was in a given image. In Task 1, this question was direct ("is the label in the image?"); in Task 2, it was indirect ("a machine predicted X, is the machine correct?"). To analyze and directly compare the combined annotations from both tasks, we converted Task 2 responses to reflect whether the human indicated that the label was in the image.

Aggregation of human scores to supermajority vote. We classify image-label pairs along three dimensions: (i) "clear yes" (positive examples) where at least 66% of annotators indicated the label was in the image, (ii) “clear no” (negative examples) where at least 66% of annotators indicated that the label was not in the image, and (iii) “ambiguous,” for all other examples that did not meet either of the previous two criteria. Image-label pairs may fall into the ambiguous category due to either a high degree of disagreement in terms of “yes”/“no” votes, or because of a high rate of “unsure” answers.

4 Results

We classify the image-label pairs from the challenge as either positive or negative examples of the submitted label. We use supermajority vote of human scores to identify which image-label pairs are positive examples (“clear yes”), negative examples (“clear no”), or could not be reliably classified due to rater disagreements or high rates of “unsure” ratings (“ambiguous”). Using the aggregated Task 1 and 2 results, we find 4300 positive examples (40.3%), 2264 negative examples (21.2%), and 4104 ambiguous examples (38.5%); Appendix table 6 breaks down these aggregate values by the target labels.

Model performance and image adversariality. As over one third of image-label pairs from the challenge were ambiguous to human raters, we investigate whether these examples were also ambiguous to machine vision models. To do this, we quantify the *adversariality of the image-label pairs* using the 61.5% of the dataset (6564 image-label pairs) on which we can compute a high-agreement human label (the “clear yes” and “clear no” examples in Table 6). Adversariality is computed as the number of human-model disagreements observed across the models tested. We identify 710 (10.8%) highly adversarial image-label pairs that none of the 6 models got correct (where “correct” means “agrees with the human consensus”). This method allows us to rank the adversariality of individual images (Table 1), based on how many models made incorrect judgements. We find that 72.8% of images were adversarial to at least one of the state-of-the-art models.

Adversarial (adv.) strength: number of models fooled	0 (not adv.)	1	2	3	4	5	6 (very adv.)
N. image-label pairs	1784	1207	1426	578	472	387	710
Percent of 6554 dataset	27.2	18.4	21.7	8.8	2.7	5.9	10.8

Table 1: Image-label pair adversariality across the dataset. To accurately reflect human-model agreement patterns, we exclude items with no human supermajority vote.

Reasons for adversariality. We break down this measure of adversariality by using the qualitative labels assigned by annotators in Task 2 to identify which model error reasons are most associated with high adversariality (Table 2). We observe that visual similarity between the label and the image (e.g., the label is “bird” and the image shows a bat) is the most frequently identified reason for model errors and is most associated with highly adversarial image-label pairs, with 55% of the 710 most adversarial images falling into the category of visual similarity. Annotators also identified misleading background context and atypical depictions of the label as primary causes of model failures, covering 30% and 33% of the most adversarial images, respectively.

Adversariality	Total pairs	Ambiguous label	Artistic depiction	Quality issue	Background context	Visual similarity	Out of context	Atypical depiction	Other error reason
1	1137	7	6	1	10	45	0	15	13
2	1404	36	21	22	570	207	6	832	549
3	562	16	7	5	216	102	0	308	156
4	471	18	5	6	172	144	4	228	107
5	387	14	5	1	114	196	2	151	85
6	710	26	15	10	212	389	0	234	125
TOTAL	4671	117	59	45	1294	1083	12	1768	1035

Table 2: Total image-label pairs for which a given error reason was indicated by at least 25% of raters in the Task 2 qualitative labeling task. "Total pairs" represents the total number of image-label pairs rated in Task 2. Totals across rows may be greater than the "total pairs," as examples can have more than one error reason.

Factors in human disagreements. We investigate potential reasons for the disagreement between humans that we observed in 38% of image-label pairs (i.e., no supermajority agreement). For this, we consider the full challenge dataset, and we assess ambiguity from three perspectives: (1) *disagreements due to rater characteristics*, (2) *disagreements due to characteristics of image-label pairs*, and (3) *disagreements due to characteristics of rating task*. These three perspectives have previously been identified as relevant to understanding crowd labels and rater disagreements Aroyo and Welty (2014). We use a linear mixed effects model (see § D.1), and we compare three single-predictor models to the null model using an ANOVA. Table 8 shows that each of these three models is a significantly better fit for the data compared to the null model, indicating that rater characteristics (as indexed by locale), label name, and task framing all explain a significant amount of variance in the data. To determine whether these three factors interact with each other, we construct both an additive and an interactive model using all three predictors; the interactive model is a significantly better fit for the data compared to the additive model ($p < 0.001$).

We used *variance partitioning analysis* to identify which of the three factors (rater locale, label name, task type) had the greatest impact on raters' judgments. Variance explained by rater id was accounted for first, and then an additive model was fitted to the residuals using features from the three factors ($R^2 = 0.159$). To understand the shared and independent variance of each set of features, several submodels were fitted to these residuals. Appendix Figure 5 shows that task type ($R^2_{uniq} = 0.079$) followed by label ($R^2_{uniq} = 0.057$) and rater locale ($R^2_{uniq} = 0.010$) have the highest amount of explained unique variance, with these features' combined unique variance accounting for 91.57% of observed variance in the original additive model. Shared variance across these features did not impact raters' judgements as much as each individual factor. To understand how these factors affect to raters' judgments, we conduct qualitative analyses of disagreements within each factor.

Understanding disagreements due to rater characteristics. For both Tasks 1 and 2, we investigate rater agreement with Krippendorf's alpha (inter-rater reliability; IRR) and cross-replication reliability (xRR). Overall agreement was only moderate in both tasks.In

Metric	Rater locale	Agreement
IRR	OVERALL	0.4737
	India	0.5739
	USA	0.5739
	Canada	0.3794
xRR	India x USA	0.5429
	India & Canada	0.4653
	USA & Canada	0.5088

Table 3: Task 1 IRR & xRR scores, by locale.

Metric	Rater locale	Agreement
IRR	OVERALL	0.1982
	India	0.3624
	USA	0.1299
xRR	India & USA	0.1846

Table 4: Task 2 IRR & xRR scores by locale.

Task 1, IRR was higher within locale for US and Indian raters than the overall IRR; xRR revealed that the Indian and American raters agreed with each other more than did Indian & Canadian raters or Canadian & American raters (Table 3). In Task 2, agreement was even lower than in Task 1.Taken together, these results show that human labelers did not tend to agree with each other on label judgments, and that a rater's locale impacted how that rater labeled images. Appendix table 9 provides example images where different locales reached different consensus labels. In panel (a), US raters affirmed the label "bird," Canadian raters rejected the label "bird," and Indian raters unanimously indicated "unsure;" in (b), 92% of American raters affirmed that the label "bird" while 86% of Indian raters were unsure. Both examples are artistic depictions of a "bird"—they are drawings that represent a bird (or just the bird's skeleton), and the different response patterns from raters in different locales highlights the way that a person's cultural context may influence their judgments in what many would consider an *objective* labeling task.

Understanding disagreements due to the image-label pairs. To identify image-label pairs that are inherently ambiguous, we identify examples where a high number of raters responded that they were "unsure" if the label was in the image. In 2039 examples (21.5% of all image-label pairs), the "unsure" label was the most frequently selected label across Task 1 raters. Appendix Table 10 shows two illustrative examples. In the first case, where the label is "Thanksgiving," it is genuinely ambiguous whether the meal is a Thanksgiving dinner; in the second it is ambiguous whether the people wearing white coats are "physicians," as opposed to any other profession that wears a lab coat. In both cases, the label is potentially consistent with the image, but crucially disambiguating background information about the image's setting, date, or participants is unavailable to the raters. Professions and roles (two of the more inherently ambiguous labels in the challenge) can be strongly context-dependent and identification relies on cultural knowledge and assumptions about the event being depicted. However, we also observe that concrete object labels (e.g., "bird") can lead to consistent unsure annotations; for example when the image is a painting of a bird, a bird mascot for a sports team, or a whole roasted chicken, annotators disagree on or are unsure about whether the label "bird" should apply.

Understanding disagreements due to the rating task. To identify cases where the task may have affected rater judgments, we analyze examples for which the supermajority vote label on a given example *changes* between the two tasks. We observe that 35.8% of the image-label pairs switch supermajority vote labels between Tasks 1 and 2 (Appendix Table 14). Most flips involve the “ambiguous” label, indicating relatively few cases where raters truly change their vote from “yes” to “no” (or vice versa). We describe observations from these cases in Appendix K, and also randomly select an image-label pair from each of the six different kinds of label flips observed to illustrate these cases.

5 Discussion & recommendations

In this paper, we are concerned with label ambiguity in large label space models, which is typically deleterious to model performance. We identified three key factors contributing to label ambiguity: rater background, label characteristics, and task design. These factors influence whether humans tend to disagree with both model predictions and each other. We demonstrated that it is, in fact, challenging for human raters and machines to agree on label ground truth, even for relatively concrete concepts such as “bird.” We further demonstrated that the geographical location in which a human rater is situated can have an impact on their answers in a labeling task. Finally, we demonstrated that small changes to the way a labeling task is framed can have an impact on how the task is performed. Given these potential complications to performing the bedrock task of machine vision model training (assigning ground truth to images), we conclude with our recommendations as to how developers, annotation guidelines and policymakers can best address label ambiguity.

- **Take a community-driven approach to data labeling.** Make sure that the people doing the labeling are from the communities that are going to be impacted by the model deployment.
- **Assume variance, ambiguity, and subjectivity are always present in any data labeling task,** regardless of how simple it may seem. There is not, and cannot be, one singular “gold standard.” To the extent possible, identify and explore potential sources of ambiguity in any data set, and understand how these sources of ambiguity might be related to the communities impacted by the model.
- **Define and deploy metrics to measure ambiguity in data.** For example, if data is labeled in different sessions, on different interfaces, or by different pools raters, measure and track differences between data subsets. Measure and track any differences across data subset by demographic properties of the community that will be impacted by the data (e.g., geographic location, gender, age, ability).

There has been little work that provides specific recommendations for policies pertaining to large label space models. Currently, content moderation strategies recommend employing machine safety filters that comprise several safety classification models Hao et al. (2023). Although our dataset does not include safety content, our challenge shows that even for categories that are non-controversial, there is ambiguity. Thus, for more subjective labels that pertain to safety (e.g., porn, violence), these ambiguities may be amplified Homan et al. (2023), which can result in unreliable safety classifications. Adopting these recommendations will ensure that a deployed model has been contributed to by the community it serves, that possible sources of model failure are understood and tracked, and that the way the model is serving different subsets of the community is also tracked. A model deployed under these conditions is on the right track to responsibly serve its community.

Reproducibility Statement

The original challenge data has been made available at `https://github.com/google-research-datasets/cats4ml-dataset`). This dataset contains the image-label pairs collected for the challenge along with an aggregation of five human annotations for each example. Note that following DMLR guidelines, this link does not need to be anonymized for review, given the inherent challenges of anonymizing dataset work. For the study described in this paper, we collected additional human annotations not part of the original repository; those annotations will be made available after review as a supplemental dataset, along with the code for the analyses conducted in this paper (descriptive stats, task score conversions, IRR, xRR, mixed-effects modelling, variance partitioning). To accompany the additional data release, we will also include a datasheet (Gebru et al., 2018).

References

Lora Aroyo and Chris Welty. The three sides of CrowdTruth. *HCj*, 1(1), September 2014.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. URL `http://arxiv.org/abs/1803.09010`.

Susan Hao, Piyush Kumar, Sarah Laszlo, Shivani Poddar, Bhaktipriya Radharapu, and Renee Shelby. Safety and fairness for content moderation in generative models. June 2023.

Christopher M Homan, Greg Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S Taylor, and Ding Wang. Intersectionality in conversational AI safety: How bayesian multilevel models help understand diverse perceptions of safety. June 2023.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. February 2015.

Xu Ji, Andrea Vedaldi, and Joao Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019.

Sanjay Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1637–1648, New York, NY, USA, February 2016. Association for Computing Machinery.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.*, 65:101759, October 2020.

Duerig T. Alldrin N. Ferrari V. Abu-El-Haija S. Kuznetsova A. Rom H. Uijlings J. Popov S. Veit A. Belongie S. Gomes V. Gupta A. Sun C. Chechik G. Cai D. Feng Z. Narayanan

D. Murphy K Krasin, I. Openimages: A public dataset for large-scale multi-label and multi-class image classification. dataset available from https://github.com/openimages, 2017.

Ka Wong, Praveen Paritosh, and Lora Aroyo. Cross-replication reliability – an empirical approach to interpreting inter-rater reliability. June 2021.

Appendix A. Data collection protocol

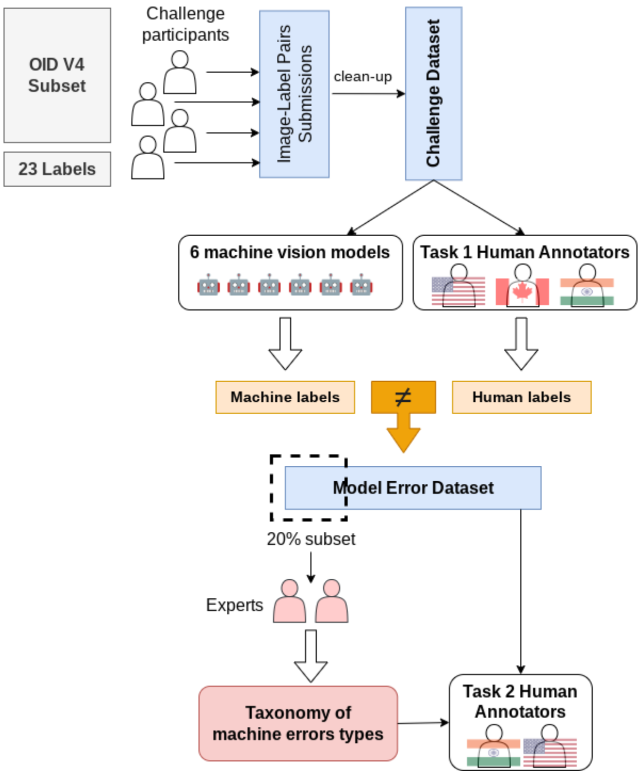


Figure 1: **Adversarial data collection from the challenge and the follow-up annotation tasks.** First, challenge participants used a subset of the OID V4 dataset to discover image-label pairs and submit them to the challenge. After cleaning the data to remove duplicates and invalid submissions, we labeled the data with state-of-the-art machine vision models and human annotators from three different locales. From their labels, we constructed a machine error dataset that consisted only of the image-label pairs with human-model disagreements. Two members of the research team qualitatively analyzed 20% of this dataset to create a *taxonomy of reasons* for the machine errors, which was then used by human annotators from two different locales to annotate the entire machine error dataset.

Appendix B. Task Interfaces

Figures 2 and 3 show the interfaces that were shown to human raters in the two annotation tasks described in the main text.

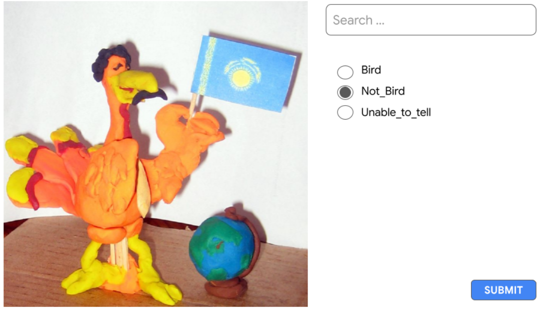
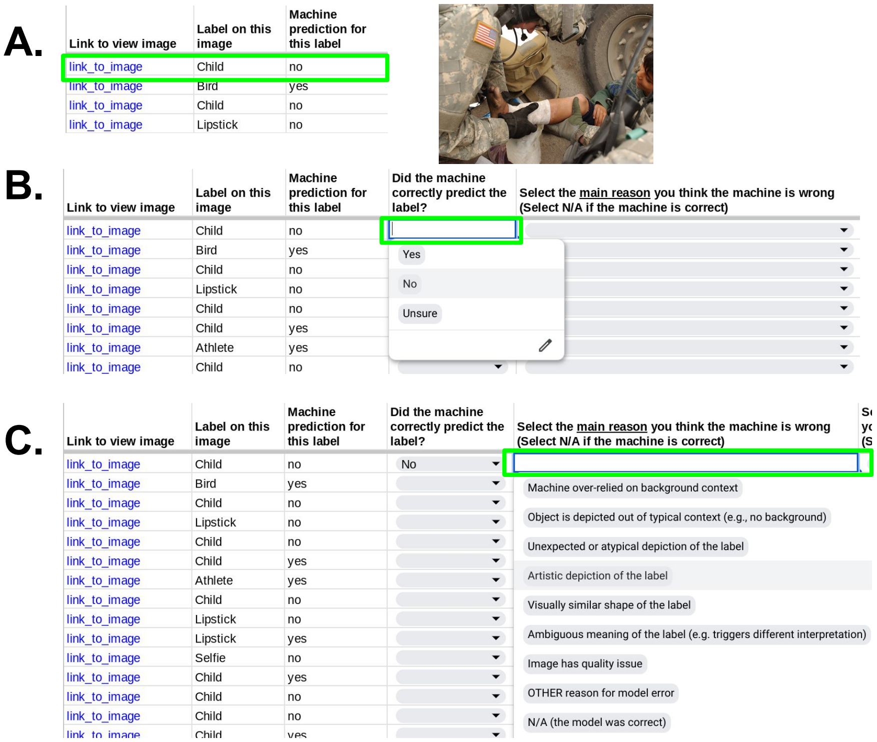


Figure 2: Sample interface for Task 1: Is label in image?


A.
B.
C.
Yes
No
Unsure
Machine over-relied on background context
Object is depicted out of typical context (e.g., no background)
Unexpected or atypical depiction of the label
Artistic depiction of the label
Visually similar shape of the label
Ambiguous meaning of the label (e.g. triggers different interpretation)
Image has quality issue
OTHER reason for model error
N/A (the model was correct)

Link to view image	Label on this image	Machine prediction for this label
link_to_image	Child	no
link_to_image	Bird	yes
link_to_image	Child	no
link_to_image	Lipstick	no

Link to view image	Label on this image	Machine prediction for this label	Did the machine correctly predict the label?	Select the main reason you think the machine is wrong (Select N/A if the machine is correct)
link_to_image	Child	no		
link_to_image	Bird	yes		
link_to_image	Child	no		
link_to_image	Lipstick	no		
link_to_image	Child	no		
link_to_image	Child	yes		
link_to_image	Athlete	yes		
link_to_image	Child	no		

Link to view image	Label on this image	Machine prediction for this label	Did the machine correctly predict the label?	Select the main reason you think the machine is wrong (Select N/A if the machine is correct)
link_to_image	Child	no	No	
link_to_image	Bird	yes		
link_to_image	Child	no		
link_to_image	Lipstick	no		
link_to_image	Child	no		
link_to_image	Child	yes		
link_to_image	Athlete	yes		
link_to_image	Child	no		
link_to_image	Lipstick	no		
link_to_image	Lipstick	yes		
link_to_image	Selfie	no		
link_to_image	Child	yes		
link_to_image	Child	no		
link_to_image	Child	no		
link_to_image	Child	yes		

Figure 3: Sample interface for Task 2: Confirm model error.

Appendix C. Label distribution in the CATS4ML dataset

In Figure 4, we show the distribution of raw counts of each label that was submitted in the CATS4ML challenge. Challenge participants were not restricted in terms of which labels they chose in their example submissions, and thus we could not ensure equal distribution across the labels. The skew towards 'bird' labels is likely due to multiple factors, including the number of instances of 'bird' in the source data, the ease of browsing images for the target object, and participant familiarity with the range of ways the label may be represented in images.

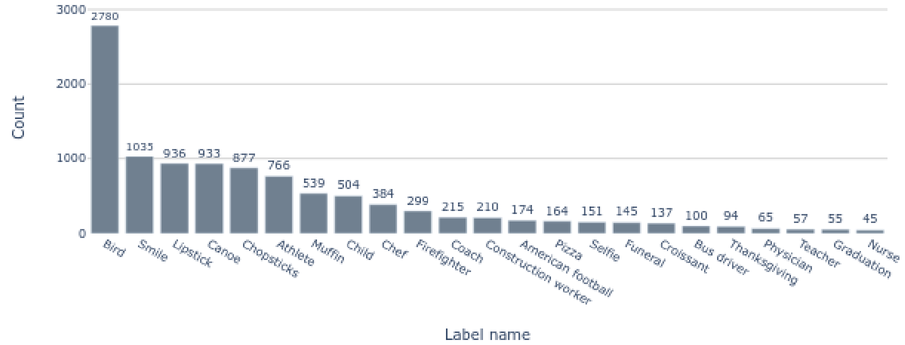


Figure 4: Histogram of valid image-label pair counts per label name.

Appendix D. Annotator details

	Size of the total rater pool				Unique raters per example			
	US raters	IN raters	CA raters	Total raters	US raters	IN raters	CA raters	Total raters
Task 1: Is label in image? *Annotated 10,668 image-label pairs*	23	13	5	41	7	7	5	19
Model error categorization *Annotated 2,035 image-label pairs*	2 experts				2 experts			
Task 2: Confirm model error *Annotated 8,326 image-label pairs*	22	34	–	56	7	7	–	14

Table 5: For each annotation task, (i) the size of rater pools, and (ii) the number of unique raters for each task example.

The annotators in Task 1 consisted of 41 raters. Table 5 (left side) shows the number of raters from each locale. We gathered 19 ratings per image-label pair (7 from raters in the US, 7 from raters in India, 5 from raters in Canada), as shown in the right side of Table 5. Each rater labeled an average of 4726 image-label pairs (with median 4088). However, 4 annotators (3 from the US, 1 from India) chose to end the task early, providing fewer than 100 annotations, so the total number of ratings provided by individual raters ranged from 3 to 9932. We ensured that each image-label pair was rated by the same number of unique annotators from the same locale distributions to ensure that the image-label-pair-level ratings were not imbalanced. Task 1 raters were compensated monetarily in alignment with local norms of the region in which they were working.

Subsequently, two members of the research team performed a qualitative analysis (See Data Analysis section) to classify the causes of model error in a sample of about 20% (2,035 image-label pairs) of the dataset from Task 1. This validation was performed in order to identify possible model error types (detailed in Appendix Table 7), and qualitatively categorize them for Task 2, described next. The experts each had in-depth experience with machine vision models.

The annotators in Task 2 consisted of 56 raters from two different locales: US and India. Tables 5 and **??** show the number of raters from each locale. As in Task 1, example-level annotations were balanced across the locales, as we gathered 14 annotations per image label pair (7 from raters in the US, 7 from raters in India). Each annotator labeled an average of 2080 image-label pairs (median 1652), with the total number of ratings provided by raters ranging from 368 to 8325. Task 2 annotators were compensated monetarily in alignment with local norms of the region in which they were working.

D.1 Data analysis

Annotator agreement metrics: We measure both inter-rater reliability (IRR, Krippendorf's alpha) and cross-replication reliability (xRR; Wong et al., 2021) to assess the agreement patterns of annotators. We measure Krippendorf's alpha because this metric is robust to imbalanced data, where different sets of annotators rate different sets of examples. Higher values of alpha indicate greater agreement among annotators. xRR is based on Cohen's Kappa, and is used to compare different groups of annotators to determine if the agreement between the two annotation distributions is more similar than would be expected by chance. xRR values are interpreted on the same scale as IRR, and higher values indicate greater similarity in responses across the two groups.

Linear modeling: Linear mixed effects models can be used to simultaneously account for random effects related to individual annotators and items, while also taking into account complex interactions between experimental conditions. We construct a null model predicting whether the rater indicated that the label is in the image or not (i.e., "yes" or "not yes", which collapses together "no" and "unsure" ratings), with random intercepts for raters and items. We compare this null model to three single-predictor models that add fixed effects of (i) rater locale, (ii) label id, and (iii) task type, and also two models that consider all three fixed effects as (i) additive, (ii) interactive predictors, and we perform model comparisons using ANOVA to compare the three single-predictor models to the null model, and to compare the additive and interactive models to ensure that we are making matched comparisons.

Qualitative analysis: Two members of the research team provided expert annotations for a qualitative analysis of the reasons for model errors. They assessed a 20% sample of the model error dataset, visually comparing the image and the model predictions for the target label on that image. The two experts proposed a taxonomy of error reasons that were then discussed with the larger research team and adapted to be used by human annotators in Task 2 to label a larger dataset. We provide examples of each error reason, with images labeled as that reason, in Appendix Table 7.

Appendix E. By-label supermajority vote results

In Table 6, we show how many images from the challenge were assigned each label ('yes,' indicating the label is in the image, or 'no,' indicating the label is not in the image), and how many were classified as 'ambiguous,' indicating that neither the 'yes' or 'no' supermajority vote label could be applied.

Target Label	clear yes	clear no	ambiguous	TOTAL
Bird	1305	43	1433	2781
Smile	721	53	261	1035
Lipstick	451	20	465	936
Canoe	63	488	382	933
Chopsticks	108	702	67	877
Athlete	630	14	123	767
Muffin	19	428	92	539
Child	387	29	88	504
Chef	32	214	138	384
Firefighter	69	70	160	299
Coach	9	19	187	215
Construction worker	49	60	101	210
American football	65	27	82	174
Pizza	87	12	65	164
Selfie	49	12	91	152
Funeral	24	23	98	145
Croissant	88	8	41	137
Bus driver	30	20	50	100
Thanksgiving	9	7	78	94
Physician	24	6	35	65
Teacher	13	3	41	57
Graduation	49	3	3	55
Nurse	19	3	23	45
TOTAL	**4300**	**2264**	**4104**	**10668**

Table 6: Counts of how many image-label pairs for each label fell into each supermajority vote category based on aggregated labels from raters in Tasks 1 and 2.

Appendix F. Qualitative labels of model error reasons

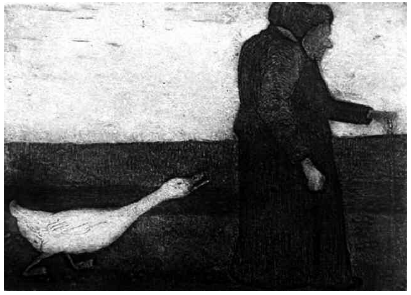
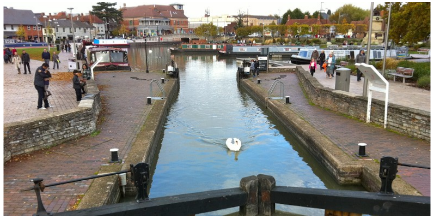
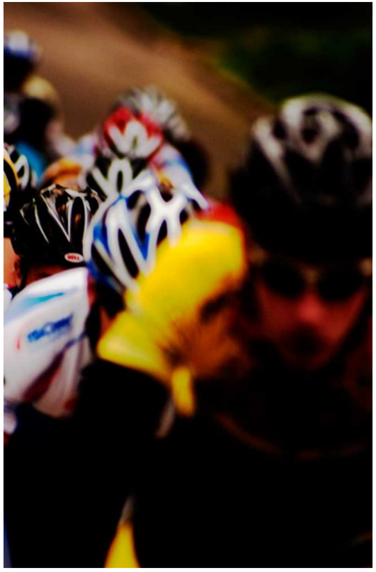
Table 7 (spanning three pages to ensure the images are legible) shows an example of each of the qualitiative labels used in the Task 2 ("confirm model error"). These labels are derived from expert validation of human-model disagreements from Task 1 ("is label in image").

Appendix G. Mixed effect model

Appendix H. Variance partitioning results

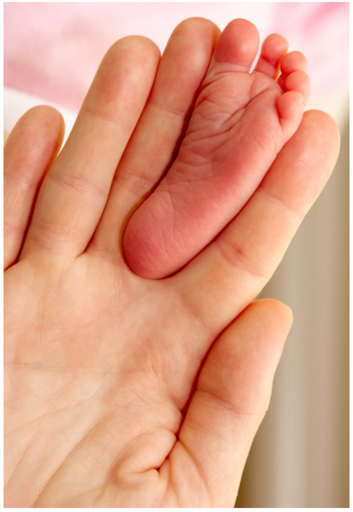
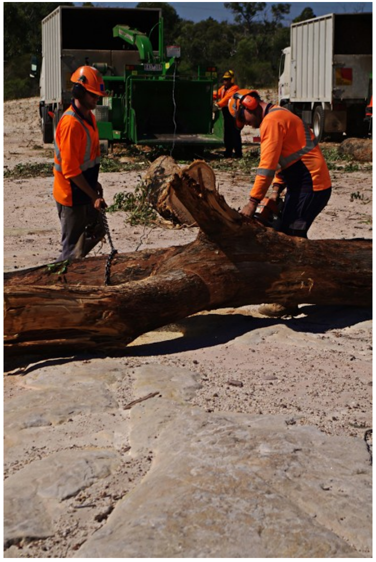
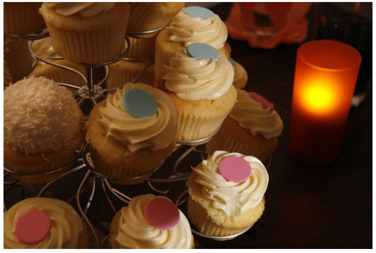
Appendix I. Examples of disagreements due to the rater locale

Table 9 shows randomly selected examples where raters from different locales gave systematically different ratings on the same image-label pair.

Error reason	Supermajority vote	Task 2 machine label	Percent of raters	Image
Artistic depiction of the label	Task 1: Ambiguous Task 2: Yes	No	78.6	**Label: BIRD** 
Machine over-relied on background context	Task 1: Ambiguous Task 2: Yes	No	85.7	**Label: BIRD** 
Object is depicted out of typical context (e.g., no background)	Task 1: Yes Task 2: Yes	No	35.7	**Label: ATHLETE** 

Appendix J. Examples of disagreements due to the image-label pair

Table 10 shows randomly selected examples where raters consistently indicated that the image itself was ambiguous with respect to the target label.

Error reason	Supermajority vote	Task 2 machine label	Percent of raters	Image
Unexpected or atypical depiction of the label	Task 1: Ambiguous Task 2: Ambiguous	No	71.4	**Label: CHILD** 
Ambiguous meaning of the label (e.g. triggers different interpretation)	Task 1: Ambiguous Task 2: No	Yes	35.7	**Label: CONSTRUCTION WORKER** 
Visually similar shape of the label	Task 1: No Task 2: No	Yes	85.7	**Label: MUFFIN** 

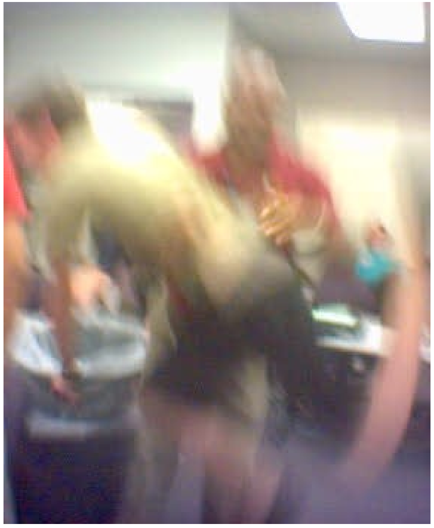
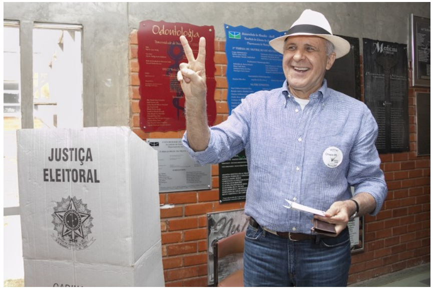
Error reason	Supermajority vote	Task 2 machine label	Percent of raters	Image
Image has quality issue	Task 1: No Task 2: No	Yes	64.3	**Label: SELFIE** 
OTHER reason for model error	Task 1: Yes Task 2: Yes	No	64.3	**Label: SMILE** 

Table 7: All error reasons from Task 2. Percent of raters indicates the percentage of Task 2 raters who indicated that the model was wrong for that particular error reason, either as the primary or secondary reason for the model error.

Model description	Model definition	AIC	BIC	Fit compared to null model
Null (baseline)	$Rating \sim 1 + (1 \mid rater\_id) + (1 \mid item\_id)$	289711.9	289754.4	N/A
Rater locale	$Rating \sim Locale + (1 \mid rater\_id) + (1 \mid item\_id)$	289677.0	289740.7	$p < 0.001$
Task type	$Rating \sim Task\_type + (1 \mid rater\_id) + (1 \mid item\_id)$	289669.7	289722.8	$p < 0.001$
Label name	$Rating \sim Label\_name + (1 \mid rater\_id) + (1 \mid item\_id)$	282069.8	282324.5	$p < 0.001$
Additive model (all predictors)	$Rating \sim Locale + Label\_name + Task\_type + (1 \mid rater\_id) + (1 \mid item\_id)$	282007.2	282293.8	$p < 0.001$
Interactive model (all predictors)	$Rating \sim Locale * Label\_name * Task\_type + (1 \mid rater\_id) + (1 \mid item\_id)$	271579.6	272725.9	$p < 0.001$

Table 8: Mixed effect model definitions and fit statistics.

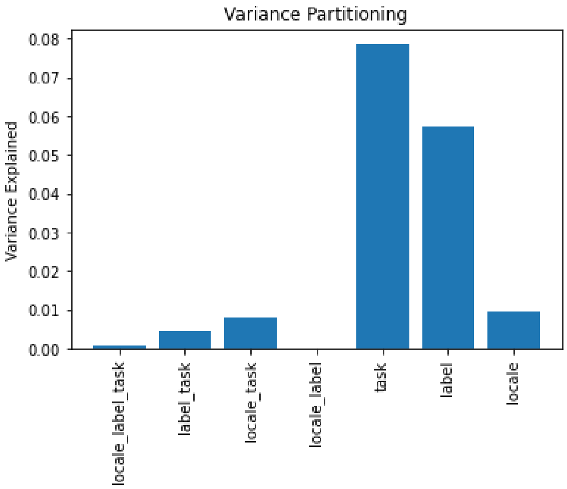
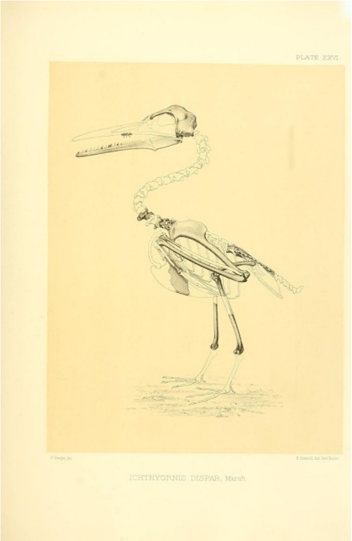


Figure 5: **Variance partitioning on a linear additive model.** First, rater id was regressed out by fitting these features to a multi-class logistic regression model with l2 penalty with raters' judgments (yes, no, unsure) as the dependent variable. Using log loss as the unit deviance or residuals, we then fit several additive models on those residuals using a combination of locale, label, and task features as independent variables. The figure above shows the shared and unique variance of these different submodels. We observe that the submodels with task followed by label and locale have the highest unique variance.

Appendix K. Examples of disagreements due to the rating task

As reported in the main text, one third of image-label pairs flip their label based on the task phrasing. Most of these flips involve the 'ambiguous' supermajority vote label, indicating that there are relatively few cases where raters truly change their vote from "yes" to "no" (or vice versa). To illustrate these cases, we randomly select an image-label pair from each of the six different kinds of label flips observed, and show the examples along with the raters' labeling patterns in Tables 11, 12 and 13. We observe patterns where the human supermajority vote label switches to align with the machine label shown in Task 2 (11a, 12a, 13b) and to contradict the machine label shown (13a). These images are illustrative of the kinds of difficulties that annotators had in assigning labels, and they show that slight changes in the wording or presentation of the task can lead to different results, even on a task that appears straightforward.

Appendix L. Task-related disagreements results

<table>
<tr><td></td><td colspan="3">A)


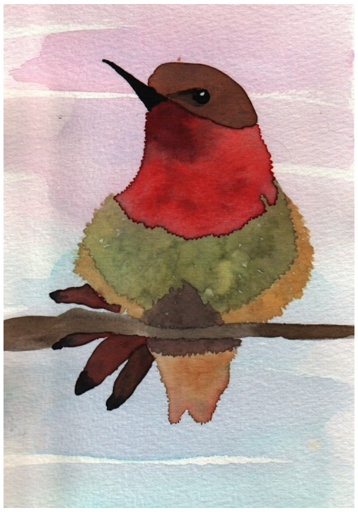
<b>Label</b>: BIRD
<b>Human majority</b>: Unsure</td><td colspan="3">B)

<b>Label</b>: BIRD
<b>Human majority</b>: Yes</td></tr>
<tr><td></td><td>Yes %</td><td>Unsure %</td><td>No %</td><td>Yes %</td><td>Unsure %</td><td>No %</td></tr>
<tr><td>US raters</td><td><b>67</b></td><td>17</td><td>17</td><td><b>92</b></td><td>0</td><td>8</td></tr>
<tr><td>CA raters</td><td>20</td><td>20</td><td><b>60</b></td><td>40</td><td><b>60</b></td><td>0</td></tr>
<tr><td>IN raters</td><td>0</td><td><b>100</b></td><td>0</td><td>0</td><td><b>86</b></td><td>14</td></tr>
</table>

Table 9: Examples of images where the raters in different locales respond differently when asked if the label is in the image.

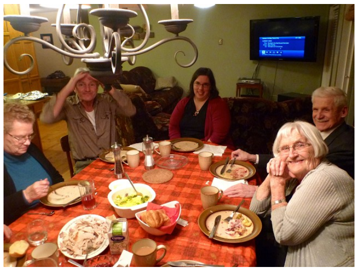
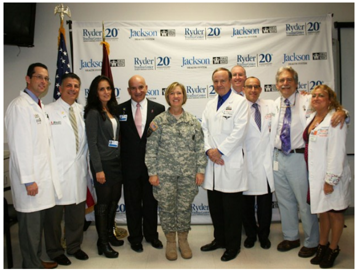
<table>
<tr><td></td><td colspan="3">A)

<b>Label</b>: THANKSGIVING
<b>Human majority</b>: Unsure</td><td colspan="3">B)

<b>Label</b>: PHYSICIAN
<b>Human majority</b>: Unsure</td></tr>
<tr><td></td><td>Yes %</td><td>Unsure %</td><td>No %</td><td>Yes %</td><td>Unsure %</td><td>No %</td></tr>
<tr><td>US raters</td><td>42</td><td><b>50</b></td><td>8</td><td>25</td><td><b>50</b></td><td>25</td></tr>
<tr><td>CA raters</td><td>40</td><td><b>60</b></td><td>0</td><td>20</td><td><b>60</b></td><td>20</td></tr>
<tr><td>IN raters</td><td>25</td><td><b>75</b></td><td>0</td><td>29</td><td><b>57</b></td><td>14</td></tr>
</table>

Table 10: Examples of images where the majority of humans indicate they are UNSURE if the label is in the image.

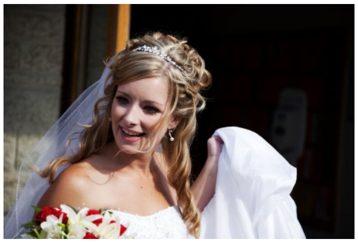
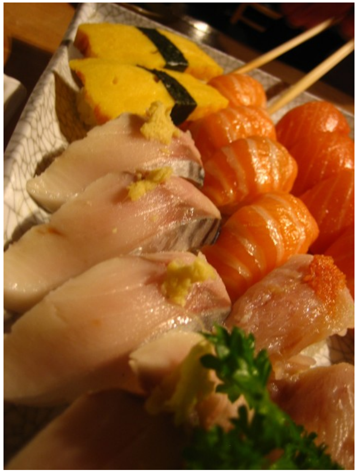
	Task 1 human label: Yes					
	Task 2 human label: No			**Task 2 human label**: Ambiguous		
	A)  **Label**: LIPSTICK **Task 2 machine label**: No			B)  **Label**: CHOPSTICKS **Task 2 machine label**: No		
	Yes %	Unsure %	No %	Yes %	Unsure %	No %
Task 1: Label in image?	**68.4**	0.0	31.6	**84.2**	10.5	5.3
Task 2: Is model correct?	28.6	0.0	**71.4**	64.3	0.0	35.7

Table 11: Examples of images where the supermajority vote label was different between the two tasks, focusing on examples that flipped an original 'yes' label in the Label-in-Image task.

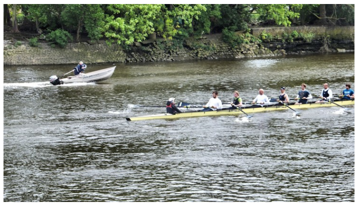
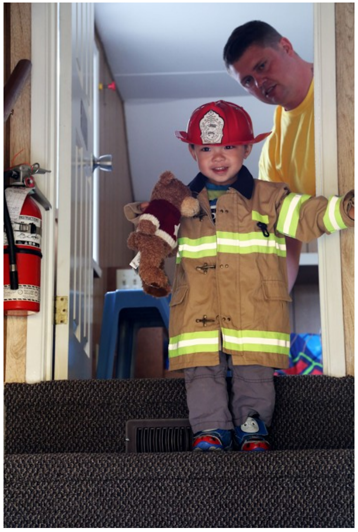
	Task 1 human label: No					
	Task 2 human label: Yes			**Task 2 human label**: Ambiguous		
	A)  **Label**: CANOE **Task 2 machine label**: Yes			B)  **Label**: FIREFIGHTER **Task 2 machine label**: Yes		
	Yes %	Unsure %	No %	Yes %	Unsure %	No %
Task 1: Label in image?	15.8	10.5	**73.7**	15.8	15.8	**68.4**
Task 2: Is model correct?	**71.4**	7.1	21.4	50.0	14.2	35.7

Table 12: Examples of images where the supermajority vote label was different between the two tasks, focusing on examples that flipped an original ‘no’ label in the label-in-image task.

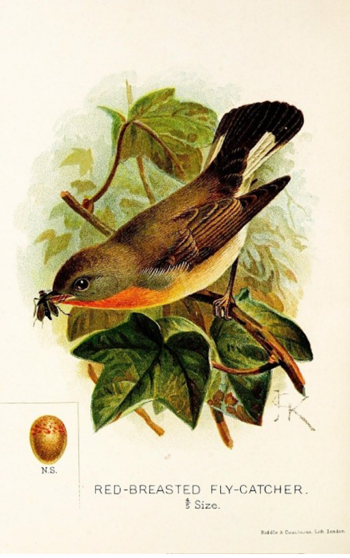
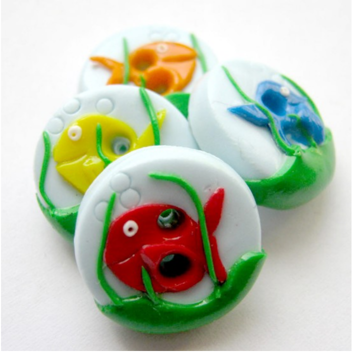
	Task 1 human label: Ambiguous					
	Task 2 human label: Yes			**Task 2 human label**: No		
	A)  N.S. RED-BREASTED FLY-CATCHER. $\frac{4}{5}$ Size. Riddle & Couchman Lith. London. **Label**: BIRD **Task 2 machine label**: No			B)  **Label**: SMILE **Task 2 machine label**: No		
	Yes %	Unsure %	No %	Yes %	Unsure %	No %
Task 1: Label in image?	31.6	63.2	5.3	0.0	52.6	47.6
Task 2: Is model correct?	**92.9**	0.0	7.1	7.1	0.0	**92.9**

Table 13: Examples of images where the supermajority vote label was different between the two tasks, focusing on examples that flipped an original ‘ambiguous’ label in the label-in-image task.

Supermajority vote label		Number of examples	Percent of total
Task 1: Is label in image?	Task 2: Is machine correct?		
Yes	**Yes**	**2714**	**32.6**
Yes	No	6	0.1
Yes	Ambiguous	464	5.8
No	Yes	9	0.1
No	**No**	**845**	**10.2**
No	Ambiguous	614	7.4
Ambiguous	Yes	1561	18.8
Ambiguous	No	325	3.9
Ambiguous	**Ambiguous**	**1787**	**21.5**

Table 14: Cross Task comparison. In **bold** are rows representing image-label pairs that had consistent supermajority labels across tasks. All other rows represent image-label pairs that had inconsistent supermajority labels across tasks.