SUPPLEMENTAL MATERIALS

MACHINE LEARNING ACCURACY

High confidence scores represent 69.5% of a test dataset. As accuracy alone is not a meaningful metric for rare attributes, for the additional attributes, we report both accuracy and recall (i.e., the proportion of such attributes correctly identified by the model; Table 1). Extremely low recall values for 'interacting' and 'young presence' are due to the rare occurrences of 'yes' annotations and, consequentially, the model's inability to learn these attributes. Count estimates by the model were generally within ± 1 count category of the true value, with a tendency to undercount rather than overcount (Table 2). The results resemble the pattern reported in Norouzzadeh et al. (2018).

Attribute	Accuracy	Recall
Moving (yes/no)	82.0%	97.3%
Eating (yes/no)	88.2%	96.6%
Standing (yes/no)	79.6%	98.1%
Interacting (yes/no)	99.0%	0.0%
Resting (yes/no)	96.7%	84.9%
Young presence (yes/no)	97.3%	5.5%

Table 1. Accuracy (predictive ability of the model) and recall (proportion of affirmative attributes correctly identified by the model) for the *Species* model on wildlife attributes of the test dataset.

Count category (true)	No. capture events (% of total)	Accuracy (exact)	Accuracy (± 1 bin)
1	16,867 (46.5%)	94.7	97.6
2	5,599 (15.4%)	52.4	95.1
3	3,350 (9.2%)	38.9	77.6
4	2,174 (6.0%)	28.2	69.5
5	1,605 (4.4%)	23.1	55.4
6	1,140 (3.1%)	13.7	42.7
7	815 (2.2%)	7.7	28.3
8	633 (1.7%)	4.9	15.6
9	394 (1.1%)	0.3	6.6
10	460 (1.3%)	0.0	72.6
11-50	3,098 (8.5%)	89.0	90.5
51+	149 (0.4%)	47.0	95.3

Table 2. Accuracy for count predictions. Reported are accuracies for exact matches between predicted and true count category (exact) and if predicted count deviated by less than one category (± 1 bin).

The out-of-sample overall performance (i.e., correct species, empty identifications) ranged from between 80.0-89.8% and rose to 92.4-98.0% for high-confidence predictions. More detailed evaluation results are available ('*Data Availability'* [2b]).

CAESAR RULES for CROWD AI

***Empty or Not* Workflow:**

For example, a capture event marked as 'Not Empty' (i.e., containing animals) by the AI with a confidence level of > 50% would be subject to the following rules:

Rule 1: if majority human votes ("human consensus") = 'Not Empty' AND number of human votes ≥ 2 AND agreement among humans = 100% AND AI confidence that 'Not Empty' > 50%, THEN remove from *Empty or Not* workflow and move to *Species* workflow.

--OR--

Rule 2: if mean human consensus = 'Not Empty' AND number of human votes ≥ 5 AND agreement among humans = 100%, THEN remove from *Empty or Not* workflow and move to *Species* workflow.

--OR—

Rule 3: if mean human consensus = 'Not Empty' AND number of human votes ≥ 10, THEN remove from *Empty or Not* workflow and move to *Species* workflow.

Inverse rules are applied to captures marked 'Empty' by the AI, which are retired from circulation without further annotation. If the first two humans viewing an image disagree with the AI designation, it is circulated for five volunteers (if 100% agreement) or ten total volunteers (if not 100% agreement) to reach a human consensus and the image is either retired or moved to the *Species* workflow.

These rules allow researchers to evaluate the level of agreement on a particular label during post-classification analyses. To minimize the risk of Type I errors, at least two citizen scientists verify every capture annotated by the AI.'. If either human classifier disagrees with the algorithm's Empty or Not designation, the capture will remain online until either five human annotations agree that it is empty (or not) or until the capture receives ten total volunteer annotations, at which point the majority vote is accepted. Note that the first retirement rule allows captures to be retired based on the AI prediction and two human confirmations, but Rules 2 and 3 in the *Empty or Not* workflow are based on volunteer consensus independent of the AI prediction.

***Species* Workflow:**

Capture events are retired when the consensus agreement among humans is 50% at 10 votes or 25% at 15 votes. If no consensus is reached, the capture event is circulated to 20 volunteers and retired with the label of the majority consensus. For specific instances, however, logical rules are in place that facilitate the classification process:

Common species: If a capture event marked as containing a common species (such as a wildebeest) by the AI with a confidence level of ≥ 90% would be subject to the following rules:

Rule 1: if human consensus = 'wildebeest' AND number of human votes ≥ 2 AND agreement among humans = 100% AND AI confidence that 'wildebeest' > 90%, THEN retire with classification 'wildebeest'.

--OR--

Rule 2: if human consensus = ‘wildebeest’ AND number of human votes ≥ 5 AND agreement among humans = 100%, THEN retire with classification ‘wildebeest’.

--OR—

Rule 3: if human consensus = ‘wildebeest’ AND number of human votes ≥ 10, THEN retire with classification ‘wildebeest’.

Sensitive species: If a capture event marked as containing a sensitive species (such as a black or white rhinoceros) by the AI with a confidence level of ≥ 20% would be subject to the following rule:

Rule 1: if human consensus = ‘rhinoceros’ AND number of human votes ≥ 1 AND agreement among humans = 100% AND AI confidence that ‘rhinoceros’ > 20%, THEN retire with classification ‘rhinoceros’.

Humans: A capture event marked as containing a human by the AI with a confidence level of ≥ 80% would be subject to the following rule:

Rule 1: if human consensus = ‘human’ AND number of human votes ≥ 2 AND agreement among humans = 100% AND AI confidence that ‘human’ ≥ 80%, THEN retire with classification ‘human’.

The threshold level of human votes needed to retire an image (confirming AI or without AI) was chosen conservatively based on simulations run on data from three of our survey sites. These simulations suggested that, based on volunteer annotation alone, only the first five volunteers were needed to reach an average accuracy of 97.33% for species identity and 90.48% for counts relative to expert-provided identifications. As such, we set five human annotations as the minimum number of votes needed to retire an image independent of AI classification and approximately half that number to retire an image if the classifications matched a highly confident AI prediction.

This species-specific functionality provides the ability to fine-tune our rules and adjust required human annotations as needed. While volunteers can easily assign labels for rare species based on distinct animal characteristics, these cannot be accurately labelled by ML alone due to small training datasets. The human consensus validates or voids the algorithm’s results and confidence scores, which helps to determine additional training goals to improve ML scores with subsequent batches of data.