Human-AI Collaboration for Conservation

Elizabeth Bondi-Kelly

MIT CSAIL
Computer Science & Artificial Intelligence Laboratory

University of Michigan
AI in Conservation

- Noisy
- Small
- Hard to collect

Humans involved in system

Park rangers
Drones
Targets

Photo Credit: Air Shepherd
AI in Conservation

- Noisy
- Small
- Hard to collect

Humans involved in system

Photo Credit: Air Shepherd
Example
Human-AI Collaboration in Practice
Human-AI Collaboration in Practice

AI Model

Use Model Prediction

Defer
Prior Work on Selective Prediction

Humans may over-rely on AI model predictions and/or may distrust an AI model

- Dietvorst, Simmons, Massey 2015
- Rastogi, Zhang, Wei, Varshney, Dhurandhar, Tomsett 2020
- Gaube, Suresh, Raue, Merritt, Berkowitz, Lermer, Coughlin, Guttag, Colak, Ghassemi 2021
- ...
Question

Does deferral still help in human-AI systems?
Deferral Model

Predictions | Ground truth | Historical human labels

AI Model | Animal | Human

Train deferral model
Training

Adding AI Assistance

We have also trained an AI model to assist humans in this task. It is correct 97% of the time.

The AI model is also equipped with deferral, so you may see a message like "AI model deferred." It means that the AI model is uncertain. On cases that are not deferred, it is correct 98% of the time.

The messages you will see communicating these results may vary. This includes cases where you may not see any message, as we are trying to collect human labels as well.

Please pay attention to the messages that are provided to you that give you context about the AI.
Selective Prediction Messaging

User would see one of the 4 conditions shown here:

- **Image 1**
- **Image 40**: AI model deferred.
- **Image 6**: AI model predicts no animal present.
- **Image 37**: AI model deferred, but predicts no animal present.

[Definitely no animal present] [Definitely animal present]
Results
Results
THANK YOU!

Elizabeth Bondi-Kelly
Email: ecbk@mit.edu / ecbk@umich.edu
Twitter: @BondiElizabeth
https://sites.google.com/view/elizabethbondi/

www.try-ai.org
Twitter: @TryAIOrg
Experiment

- 198 Prolific participants recruited
- 80 deferred images (20 TN, 20 TP, 20 FP, 20 FN)
  - No repetition
  - No message that all are deferred
  - Randomly allocated across 4 conditions
- 4 separately seeded random allocations (about 50 participants each)
Results