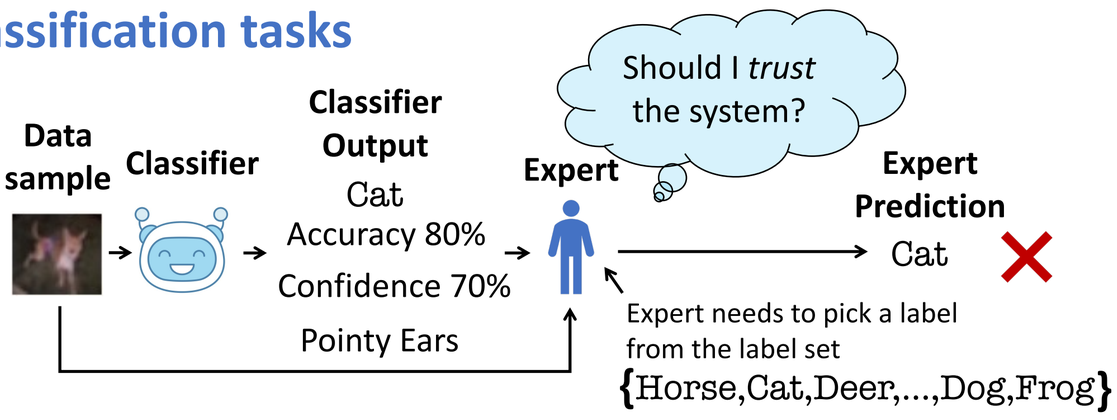
Designing Decision Support Systems with Counterfactual Prediction Sets

Max Planck Institute for Software Systems

Eleni Straitouri and Manuel Gomez-Rodriguez

Existing decision support systems for multi-class classification tasks



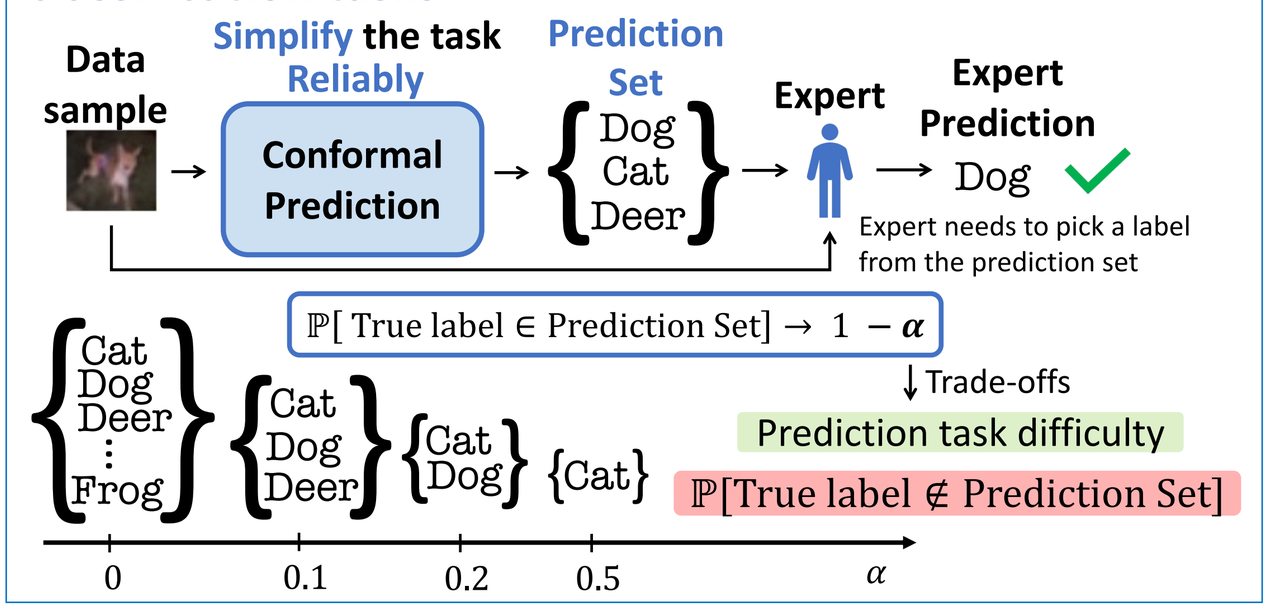
⇨ Experts ***have to decide when*** to trust the system.

⇨ Empirical findings [1,2] have reached ***no conclusion*** on how experts can avoid developing a ***misplaced*** trust on the system.

Goal: Design a decision support system that does not require humans to decide *when* to trust its predictions

[1] A. Papenmeier, G. Englebienne, and C. Seifert. How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652, 2019. [2] M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. arXiv preprint arXiv:1903.12220, 2019.

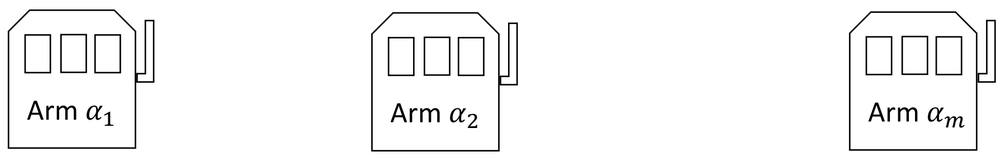
Our decision support system for multi-class classification tasks



Optimizing conformal predictors with bandits

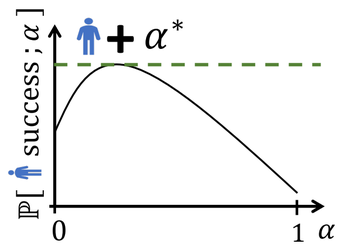
Arms ⇔ α values

$\mathbb{P}[$ success $; \alpha_1]$ $\mathbb{P}[$ success $; \alpha_2]$ ••• $\mathbb{P}[$ success $; \alpha_m]$



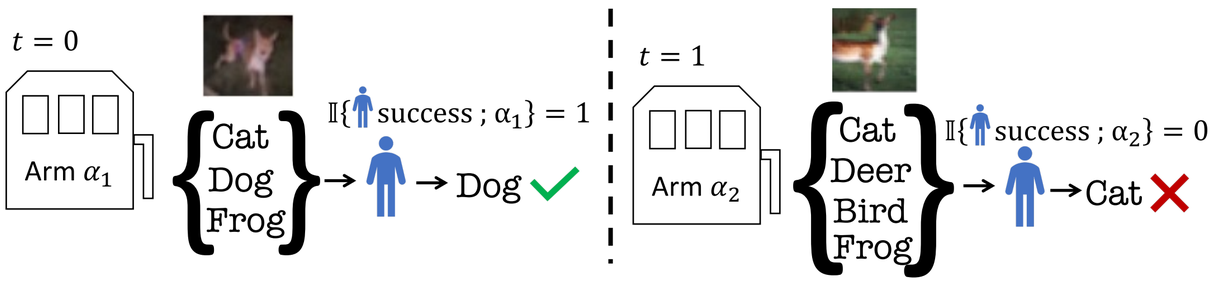
Discover optimal arm

$$\alpha^* = \underset{\alpha}{\operatorname{argmax}}\, \mathbb{P}[\text{ success } ; \alpha]$$

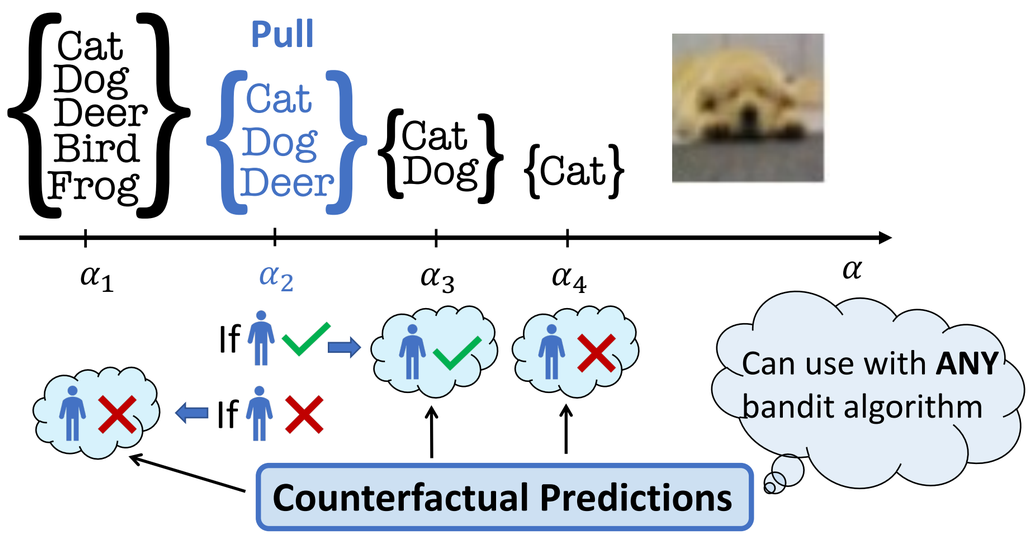


Successive Elimination

Pull arms sequentially



Counterfactual Monotonicity Assumption



Counterfactual Successive Elimination finds α^* exponentially faster

Large-scale human subject study

⇨ 194,407 predictions

⇨ 2,751 human subjects

⇨ 19,200 different pairs of natural images and sets

Example

Which one of the following categories fits better the image below?

○ Car
○ Airplane
○ Truck

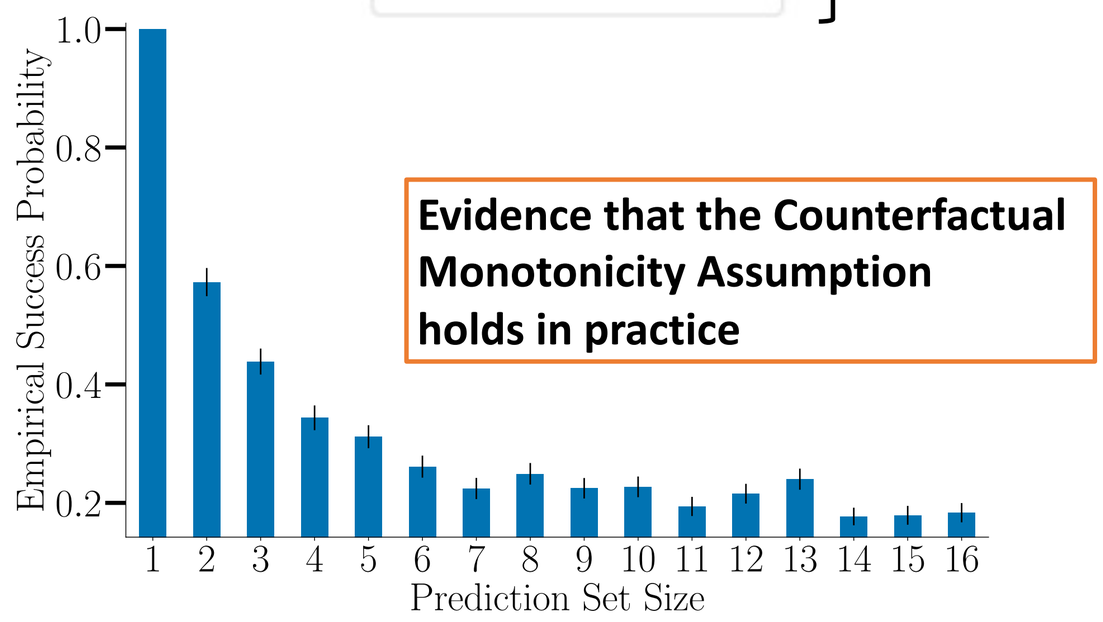
Strict implementation: picks only from the prediction set

○ Car
○ Airplane
○ Truck
○ Other

If you chose 'Other' above, please choose a category:

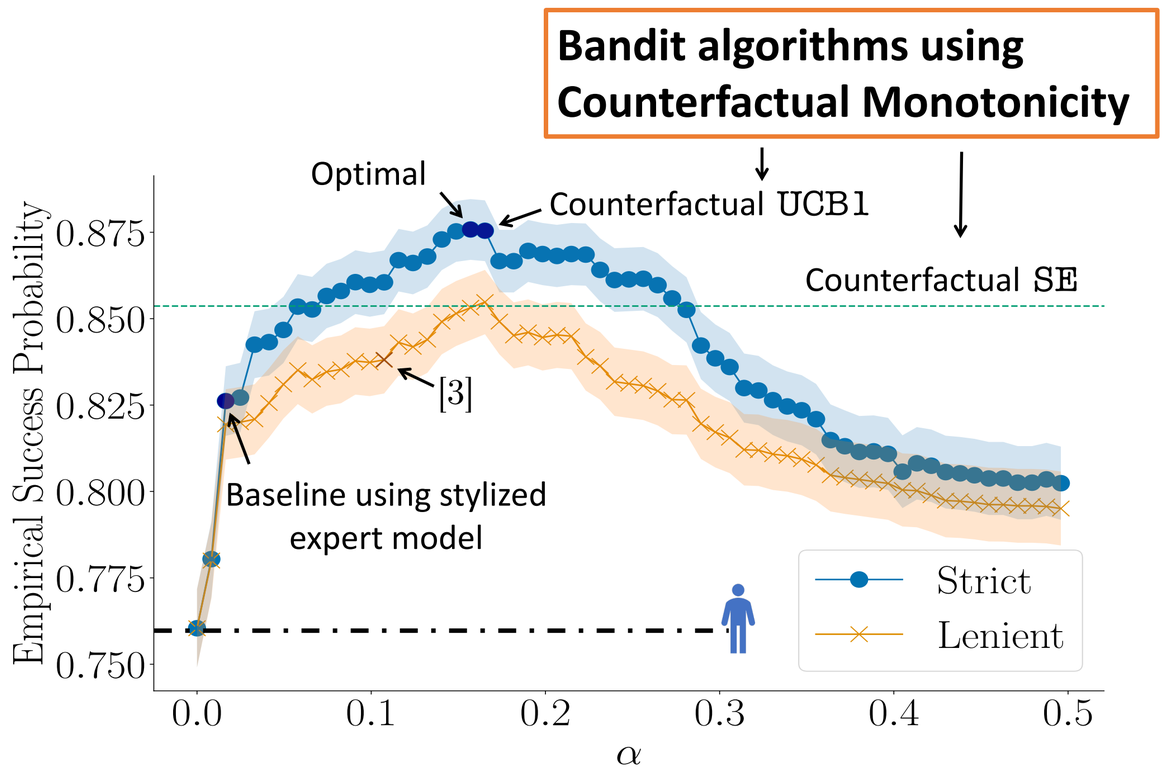
Choose

Lenient implementation: can pick any label



Evidence that the Counterfactual Monotonicity Assumption holds in practice

Bandit algorithms using Counterfactual Monotonicity



Lenient implementation

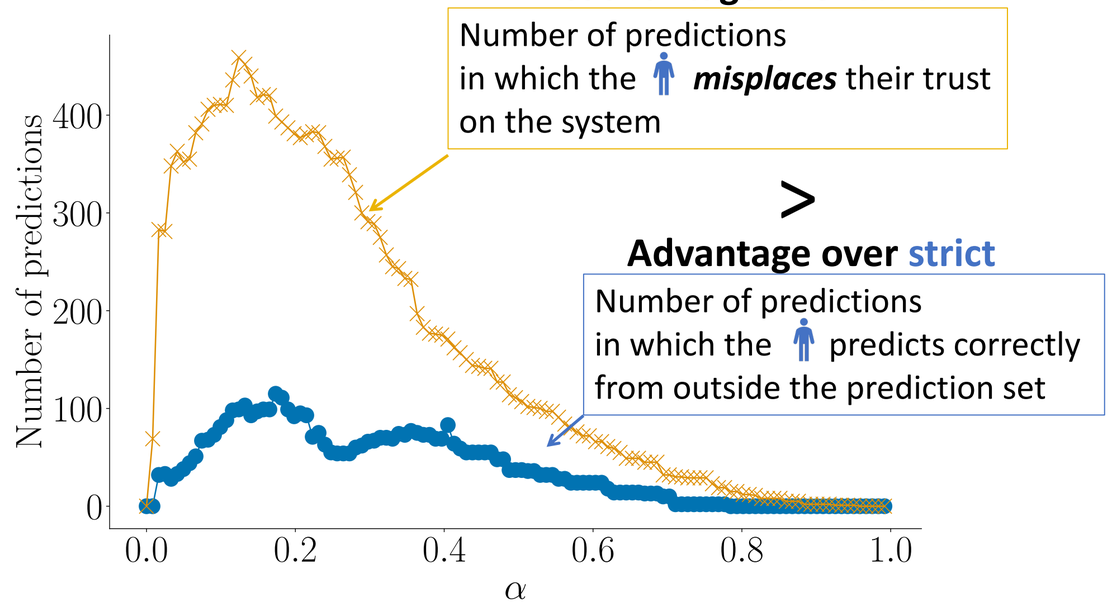
Disadvantage over strict

Number of predictions in which the ***misplaces*** their trust on the system

>

Advantage over strict

Number of predictions in which the predicts correctly from outside the prediction set



[3] Babbar, V., Bhatt, U., & Weller, A. (2022). On the utility of prediction sets in human-ai teams. arXiv preprint arXiv:2205.01411.