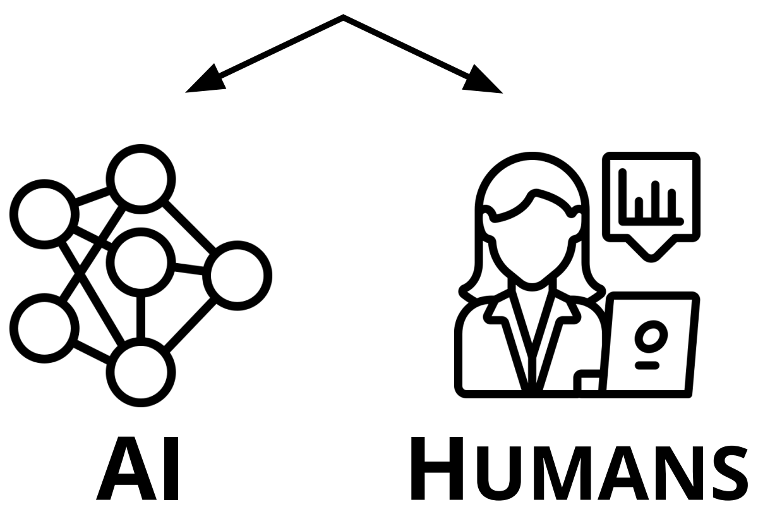
Human-AI Collaboration in Decision-Making:
Beyond Learning to Defer

Diogo Leitão | Pedro Saleiro | Mário A. T. Figueiredo | Pedro Bizarro

Human-AI Collaboration

- **AI** is now fast, scalable, and often quite accurate
- **Humans** learn fast, accrue experience, and may have access to exclusive information
- Through synergistic teaming, **human-AI collaboration** has the potential to **outperform** humans and AI in isolation
- **Key challenge**: **who should decide** in each case?



Learning to Defer

- **Confidence-based deferral:** defer to humans instances of high model uncertainty
- Madras et al. (2018): **optimal deferral depends** on model and **human performance**
- Proposed ***learning to defer***: jointly training a classifier and an assignment system to maximize **performance** (and, optionally, **fairness**)

$$\mathcal{L}_{system}(\hat{Y}_M, \hat{Y}_H, s) = \sum_i [(1 - s_i)\mathcal{L}_{CE}(Y_i, \hat{Y}_{M,i}) + s_i\mathcal{L}_{0\text{-}1}(Y_i, \hat{Y}_{H,i}) + s_i\gamma_{defer}]$$

Limitations & Challenges

Missing Human Predictions

- Learning to defer **requires human predictions for every instance** in the training set (or that the missing predictions be imputable)
- Often unfeasible: burden of decision-making may already be shared with AI
- Imputation: not valid unless the assignment system is random (rare: confidence-based deferral has substantial performance gains — Hendrycks & Gimpel, 2016)

Joint Learning

- Benefit: the main classifier can **focus** on instances humans cannot solve
- **Drawbacks**:
  1. In use-cases where the AI **advises** humans, it will be rendered useless
  2. If humans become temporarily or partially **unavailable**, the AI will be unable to substitute them (purposely not trained in those areas)

Multiple Humans & Capacity Management

- Keswani et al. (2021), Hemmer et al. (2022): extend L2D to a multiple-expert setting
- **Drawback**: now **requires human predictions from every human for every instance**
- Often **unfeasible**: in performative use-cases, having more than one human review each instance is highly inefficient
- Imputation: generalization may falter if past assignments were not random (not i.i.d.)

Selective Labels

- Originate from decision-making processes where **predictions influence outcomes**
- Ubiquitous in high-stakes environments (e.g. bail decisions, lending decisions, fraud detection)
- Learning to defer cannot deal with selective labels
- Alternative approaches require a **change of angle** or **additional assumptions**

Promoting Fairness

- Both **machine learning models and humans may be biased** against protected groups
- **Considering the specific biases** of each allows the collaboration system to mitigate unfairness
- On the contrary, introducing fairness-unaware deferral systems has been shown to aggravate unfairness
- Fairness is both an **opportunity** and a **threat**

Dynamic Environments

- **Non-stationarity factors** render ML models **obsolete** (e.g. concept drift, adversarial classification, performative prediction)
- Human-AI collaboration systems may also suffer from **change in human behavior** due to exogenous factors, or in response to the new assignment system
- Systems must be updatable with new data to keep up
- Learning to defer is not updatable as it requires human predictions for every training instance