Human-AI Collaboration

Introduction

- AI practitioners typically strive to **develop the most accurate systems**

- However, in practice, AI systems often are used to **provide advice** to people

- In such AI-advised decision making, humans and machines form a team, where the **human is responsible for making final decisions.**

- **Does high AI accuracy translate to high team performance?**

- The answer is “it is **not necessarily true**”

- Ultimately, they would want to **optimize the team performance**

Literature Review

There are four types of studies

1) Human biases
2) Methods to optimize models
3) Explainable AI design
4) Human modeling

Literature Review - Human Biases

Paper	Year/cite	Description	Results
Investigations of Performance and Bias in Human-AI Teamwork in Hiring	2022/5	How **model's predictive performance and bias may transfer to humans** in a recommendation-aided decision task. (gender bias)	- Different models have different biases - **H + DNN = unbiased, H + BOW = biased** - H+BOW was more accurate than our H+DNN, posing a trade-off between high team accuracy vs. low team bias.
To Trust or to Think: Cognitive Forcing Functions Can Reduce Over-reliance on AI	2021/81	Study on **simple explainable AI vs cognitive forcing interventions.** 1) SXAI (explanation, confidence) 2) CFF (on demand, update, wait) 3) no AI-baseline). Audit their work on people with different **Need for Cognition (high low)**	- Both improved the performance - CFF help people to **over rely significantly less** and made significantly more corect decision than SXAI - No AI is more mentally demanding than others - People perceived **SXAI to be less complex** - **CFF benefit high NFC more than low NFC** - (NFC capture motivation to engage in effortful mental activities)
Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks	2021/30	Explore heuristic people use to adjust their reliance when performance feedback is limited. (**Human-model agreement**)	- Without feedback people seem to use **level of agreement to decide how much to rely on the model** - Even after feedback, people whose **prediction is not independent and are confident,** they may still overly rely on models that share the same biases as them - **agreement shows limited** impact in influencing reliance **after people have observed the model's performance in practice.**
You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift	2021/15	How people rely on machine earning models to make decisions under **covariate shift.** (house price prediction)	- Laypeople tend to rely on an **ML model more when covariate shift occurs,** effectively resulting in over-reliance on an ML model when its performance is poor because they **expect the model to maintain its performance on out-of-distribution data**, while they believe their own decision- making performance would decrease on those data. - **Education > visualization**
Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance	2020/158	Can explanation help lead to complementary performance? (beer book reviews and LSTA)	- Showing **AI's confident achieved complementary performance**, others ExplainTop1, ExplainTop2, Adaptive no better complementary performance

Literature Review - Methods to optimize model

Paper	Year/cite	Description	Results
Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork	2021/33	**Maximize team utility** in terms of the quality of **final decision, cost of verifying and individual accuracies of people and machines**	- Expected team utility is improved though less accurate - New classifier makes more **high confidence prediction** - New classifier **sacrifices accuracy on the uncertain example to make high number of high-confidence predictions**
Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff	2019/146	Notion of "**compatibility**" of an AI update with prior user experience and present methods for studying the role of compatibility in human-AI teams. They also proposed a **re-training objective to improve the compatilbility of an update by penalizing new errors.**	- as the number of features increases, **team performance decreases because it becomes harder to create a mental model.** - as errors become more stochastic, it becomes harder to create a mental model, deteriorating team performance - a **more accurate but incompatible classifier results in lower team performance** than a less accurate but compatible classifier (no update). - compatible updates not only improve team performance but they can also r**educe the cost of retraining** users after deploying system updates
Beyond accuracy - the role of mental models in human-ai team performance	2019/180	The **human's mental model** of the AI capabilities, specifically the AI system's error boundary. Highlight two key properties of an AI's error boundary, parsimony and stochasticity, and a property of the task, dimensionality.	- Build AI systems with **parsimonious error boundaries.** - **Minimize** the **stochasticity** of system errors. - **Reduce task dimensionality** when possible either by eliminating features that are irrelevant for both machine and human reasoning or most importantly by analyzing the trade-off between the marginal gain of machine performance per added feature and the marginal loss of the accuracy of human mental models per added feature. - During model updates, deploy models whose **error boundaries are backward compatible**, i.e. by regularizing in order to minimize the introduction of new errors on instances where the user has learned to trust the system.
Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer	2018/94	Two-staged framework, automated model and an external decision-maker. Learning to defer can make systems not only more accurate but also less biased.	- Learning-to-defer achieves a **better accuracy-fairness tradeoff than rejection learning.** - It can a**daptively defer at different rates for the two sensitive groups to counteract the DM's bias**; and it is able to modulate the overall amount that it defers when the DM is biased - Most accurate model is mixture-model

Literature Review - Methods to optimize model

Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff

- Updates can arbitrarily change the AI's error boundary, introduce new errors which violate user expectations and decrease team performance.
- Locally compatible update $\forall x, [m(x) \wedge A(x, h_1(x))] \Rightarrow A(x, h_2(x))$
- Globally compatible update $\forall x, A(x, h(x)) \Rightarrow A(x, h_2(x))$
- Compatibility score $\mathcal{C}(h_1, h_2) = \frac{\sum_x A(x, h_1(x)) \cdot A(x, h_2(x))}{\sum_x A(x, h_1(x))}$

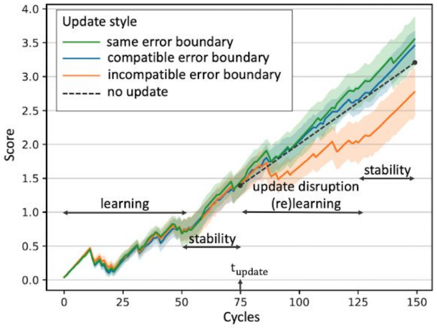
m(x): mental model of user's trust
A(x,u): whether u is an appropriate action for x
h1: learned model
h2: updated model

- Loss

$$L_c = L + \lambda_c \cdot \mathcal{D}$$

$$L(x, y, h_2) = y \cdot \log p(h_2(x)) + (1 - y) \cdot \log(1 - p(h_2(x)))$$

$$\mathcal{D}(x, y, h_1, h_2) = \mathbb{1}(h_1(x) = y) \cdot L(x, y, h_2)$$



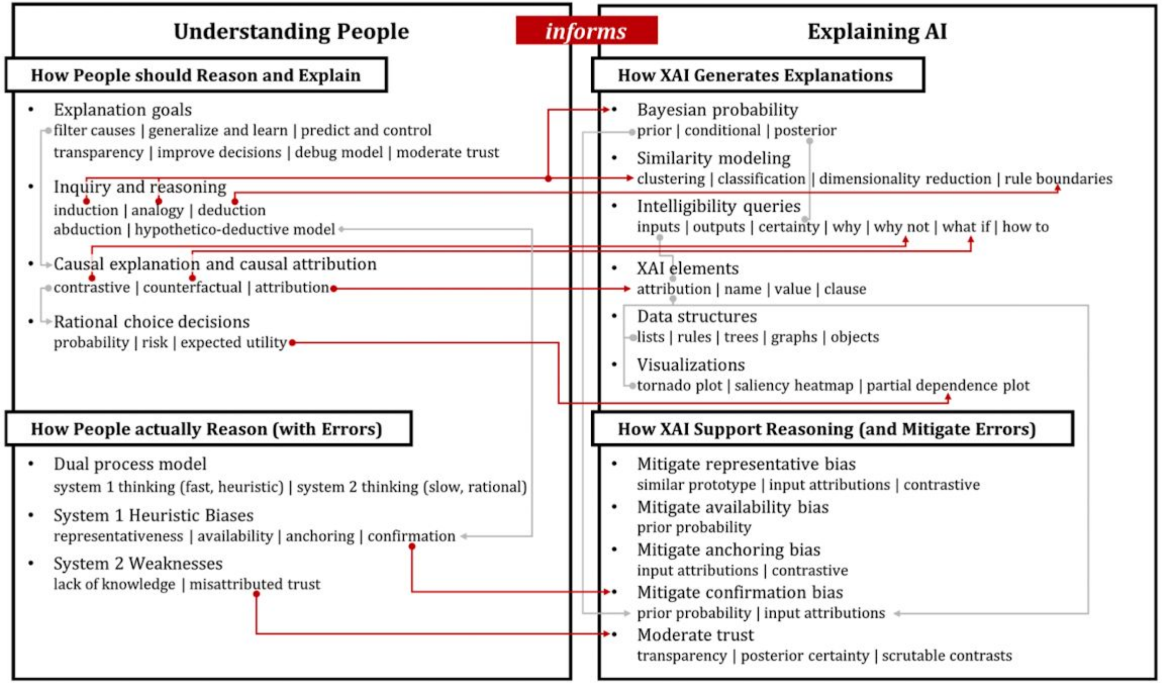
So D measures if h1 is correct and penalizing by the degree which h2 is incorrect

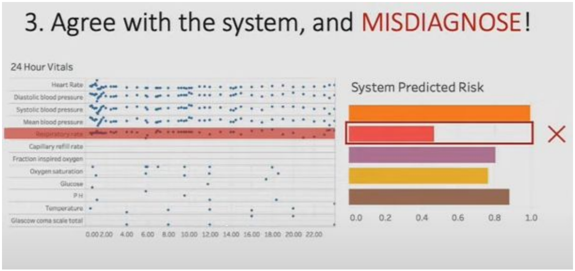
D: Dissonance
p(h2(x)): confidence in h2(x) in prediction
y: true label
1: indicator function

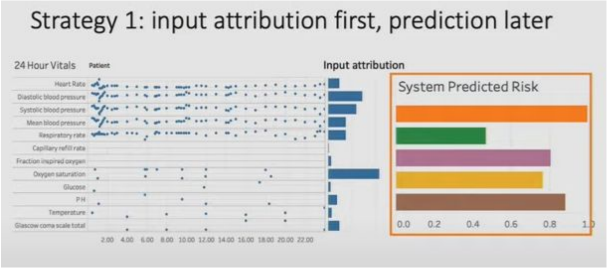
Literature Review - XAI design

Paper	Year/cite	Description	Results
Human-Centered Explainable AI (XAI): From Algorithms to User Experiences	2021/31	**Explainability and transparency are not the answer to human-ai collaboration**	- Encourage dual process - From social science 1) explanation are often contrastive, sought to response to some couterfactual cases, this is because WHY question often triggered by absonoal or unexpected eents 2) explanations are selected by the explainer, often in a biased manner. 3) Explanations are social, as a transfer knowledge, often part of a conversaion or interaction and presented relative to explainer's belief about the explainee's belief 4) probabilities or statistical information to explain is often ineffective and unsatisfying.
Are explanations helpful? a comparative study of the effects of explanations in ai-assisted	2021/61	Highlight three desirable properties that ideal AI explanations should satisfy—improve people’s understanding of the AI model, help people recognize the model uncertainty, and support people’s calibrated trust in the model.	- The **effects of model explanations are dramatically different on tasks where people have varying levels of domain expertise** - **Feature importance** explanations increase people’s **objective understanding** of an AI model, while **feature contribution** explanations increase people’s **subjective understanding** of an A - For expert people, feature contribution is most satisfying - Counterfactual helps experts but fails to help calibrate trust though closest to how human explain decisions
“Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human–AI Collaborative	2019/209	Interview doctors	- Pitfalls of AI (error rates relative to human (not necessarily to themselves)) - Assume what they consider a difficult case, AI would also struggle - Data and its generalization - Similar opinion - Subjective decision threshold - To some participant, when the AI’s objective is to be as accurate as possible, they quickly lost trust when AI makes mistake
Designing Theory-Driver User-Centric Explanable AI	2019/458	A theory-driven conceptual framework linking different XAI explanation facilities to user reasoning goals which provides pathways to mitigate reasoning failures due to cognitive biases.	Next page

Literature Review - XAI design







Literature Review - Human Modeling

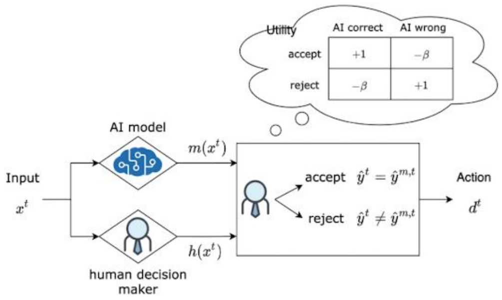
Paper	Year/cite	Description	Results
Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making	2022/20	Use knowledge from field of **cognitive science to account for cognitive biases** in human-AI collaborative decision making setting, and mitigate their negative effects on collaborative performance. They **study and provide mitigating strategies for anchoring biases** in AI-assisted decision-making	- participants' likelihood of sufficiently adjusting away from the incorrect AI prediction increased as the time allocated increased - confidence based time and confidence based time with explanation out perform AI only and human only groups
Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making	2022/2	a quantitative understanding of whether and when would human decision makers adopt the AI model's recommendation. Human decision makers' **utility evaluation and action selection** are influenced by their own judgement and confidence on the decision-making task.	- Added human adjusted selection component significantly increases the model's performance in fitting human behavior data - shows that human incorporates their own judgement in a decision-making trial to decide whether to accept an AI model's recommendation

post $$\mathbb{P}(\tilde{Y}=i|D)=\frac{\mathbb{P}(D|\tilde{Y}=i)\mathbb{P}_{pr}(\tilde{Y}=i)}{\sum_{j\in\{0,1\}}\mathbb{P}(D|\tilde{Y}=j)\mathbb{P}_{pr}(\tilde{Y}=j)},$$ prior

$$\mathbb{P}(\tilde{Y}|D,f(M))\propto\mathbb{P}(D|\tilde{Y})\mathbb{P}(f(M)|\tilde{Y})\mathbb{P}_{pr}(\tilde{Y}),$$

$$\frac{\mathbb{P}(\tilde{Y}=1|D,f(M))}{\mathbb{P}(\tilde{Y}=0|D,f(M))}=\left(\frac{\mathbb{P}(D|\tilde{Y}=1)}{\mathbb{P}(D|\tilde{Y}=0)}\right)^{\alpha}\left(\frac{\mathbb{P}(f(M)|\tilde{Y}=1)}{\mathbb{P}(f(M)|\tilde{Y}=0)}\right)^{\beta}\left(\frac{\mathbb{P}_{pr}(\tilde{Y}=1)}{\mathbb{P}_{pr}(\tilde{Y}=0)}\right)^{\gamma}$$

- *Anchoring bias: when beta is high*
- *Confirmation bias: gamma is high as the weight on data and machine prediction preduces*
- *Direction of distortion is alpha*



Naive Bayes

$$c^{m+h,t}=\frac{1}{1+\frac{(1-c^{m,t})\cdot(1-h(x^t)[y^t=\hat{y}^{m,t}])}{c^{m,t}\cdot h(x^t)[y^t=\hat{y}^{m,t}]}}$$

Adjusted Naive Bayes

$$c^{m+h,t}=\frac{1}{1+\frac{(1-a)\cdot(1-b)}{a\cdot b}},\text{where } a=\frac{(c^{m,t})^{\gamma}}{(c^{m,t})^{\gamma}+(1-c^{m,t})^{\gamma}},$$

$$b=\frac{(h(x^t)[y^t=\hat{y}^{m,t}])^{\gamma}}{(h(x^t)[y^t=\hat{y}^{m,t}])^{\gamma}+(1-h(x^t)[y^t=\hat{y}^{m,t}])^{\gamma}}$$

Summary

- There is no one size fits all solution
- AI practitioners will still optimize models in isolations
- In practise, it is not about how accurate the models are, but rather when and how to use them.
- Focus should be paid more on 1) mental model of the users 2) how to update model appropriately
- Adaptive solution is a key: explainability of the model, cognitive forcing interventions
- Human biases
- Expertise

Idea:

- According to prospect theory: since people are more effected by loss more, we can show how likely error could occur
- People are selective: Since people meet decision conclusion when they find enough information to support their decision

Problems:

- It is no longer NLP and not even in health domain
- I still don't have a 3 year plan
- It's missing "what drives your economic engine?"