Machine Learning for Healthcare
6.7930, HST.956

Lecture 17: Human-AI Collaboration in Healthcare

David Sontag

Acknowledgement: slides adopted from
Hussein Mozannar

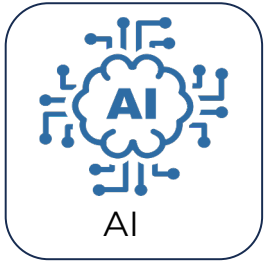
CSAIL

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE

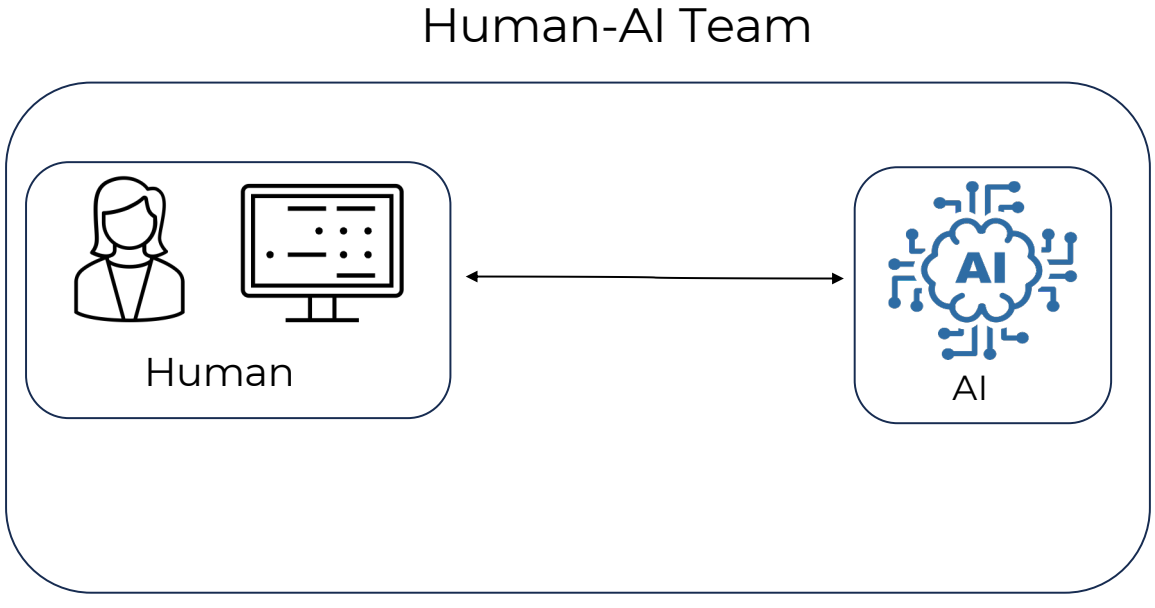
HEALTH SCIENCES & TECHNOLOGY

You’ve built an AI model, but how is it actually used?

The premise of AI is to automate tasks, but often that is not feasible, and it might not be our best option!



Often there is a human in the loop!



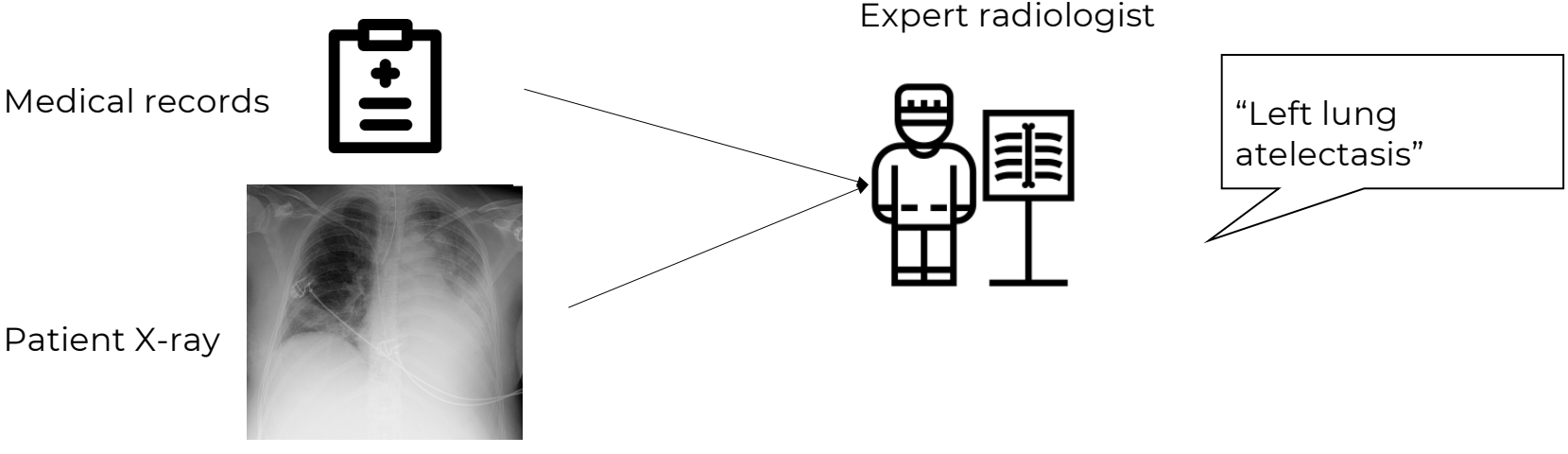
Today's Lecture

1. **How do we combine humans and AI?**
   - Modes of human-AI interaction
2. **How do people think about AI?**
   - Mental Models
3. **How do we interact with generative AI?**
   - AI-assisted reading and writing

Let’s start with an example task to solve

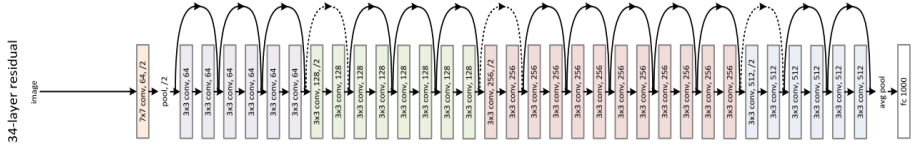
Detecting Atelectasis From Chest X-rays

- Atelectasis: the collapse of part or all of a lung.
- Can be caused by mucus, foreign objects or tumors blocking the airway.



Detecting Atelectasis From Chest X-rays

- A student from class decided to build an ML model for detecting Atelectasis instead.
- They use CheXpert [1] dataset of >200k chest x-rays with annotations
- They train a ResNet-34 model [2]



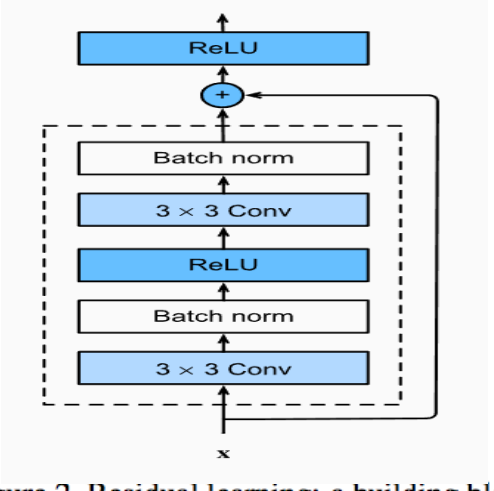
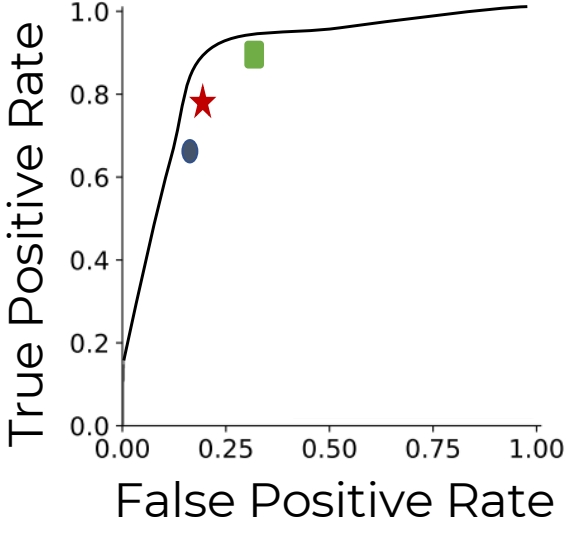


Figure 2. Residual learning: a building block.

[1]: Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. 2019. [2]: He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

AI vs Human performance

- Test set: 500 x-rays annotated each by 5 radiologists, ground truth is their majority vote. 3 other radiologists to compare to.



Model (AUC = 0.91)
Rad1 (0.21,0.80)
Rad2 (0.18,0.71)
Rad3 (0.31,0.92)

Model outperforms all 3 radiologists

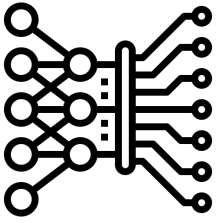
How do we integrate the AI into the current pipeline?

Deploying the AI to replace the radiologist

- **Model in isolation:** after X-ray is taken, the model makes its prediction, then referring physician can give treatment

Patient X-ray

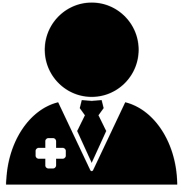
Model



Radiology Report

Heart size upper normal but stable. Mediastinal contours within normal limits. Minimal right middle lobe atelectasis. No focal airspace consolidation, pleural effusion, or pneumothorax. Degenerative endplate changes of the spine. [1]

Physician



[1]: Buendía, Félix, Joaquín Gayoso-Cabada, and José-Luis Sierra. "An Annotation Approach for Radiology Reports Linking Clinical Text and Medical Images with Instructional Purposes." Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality. 2020.

Model in isolation: Diabetic Retinopathy

- **Diabetic Retinopathy:** diabetes complication affecting the eye
- **Why we need AI:** access to care is a huge problem, especially in places like India (70mil diabetics, 2 months to get results, need to travel)

- **Model:** Dataset from Thailand, model reduces FNR by 23% but increases FPR by 2% [1]

[1]: Ruamviboonsuk, Paisan, et al. "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program." *NPJ digital medicine* 2.1 (2019): 1-9.

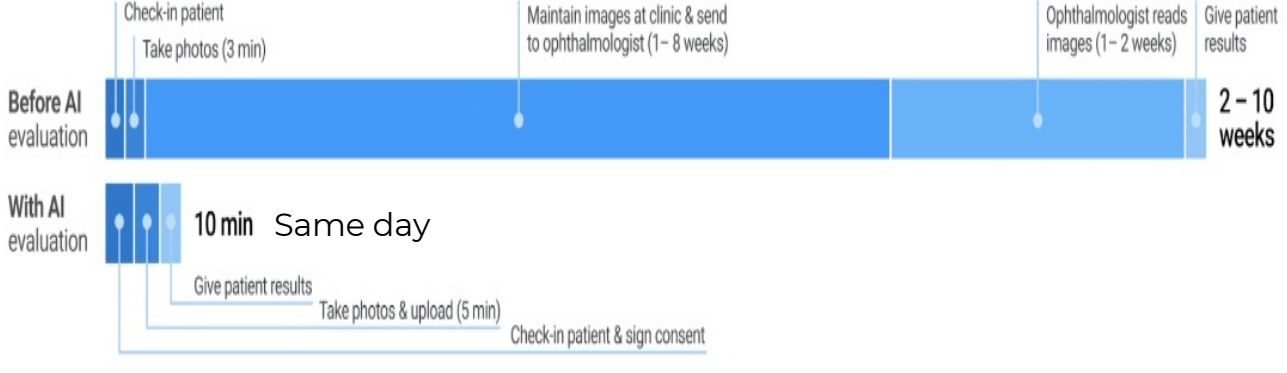
Deployment details

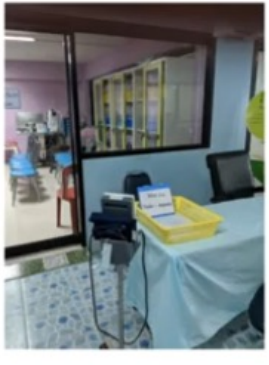
- Model deployed in 8 sites in Thailand, 1.5-year study, 7600 patients
- 200 patients/day, 5 hours wait, 90sec eye exam

[1]:Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Deployment details

- Prospective study after deployment with the nurses taking the images [1]





[1]:Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Results after deployment

- Model refused to predict on 20% of images, images were unreadable to the model
  - Imperfect lighting conditions
  - Old cameras
  - Limited time to align patients
- Nurse's observations:

"Some images are blurry, and I can still read it, but the system can't", "it's good but I think it's not as accurate. If [the eye] is a little obscured, it can't grade it"

- **Those ungraded, now needed to travel to see an ophthalmologist instead of just waiting for image to be read.**

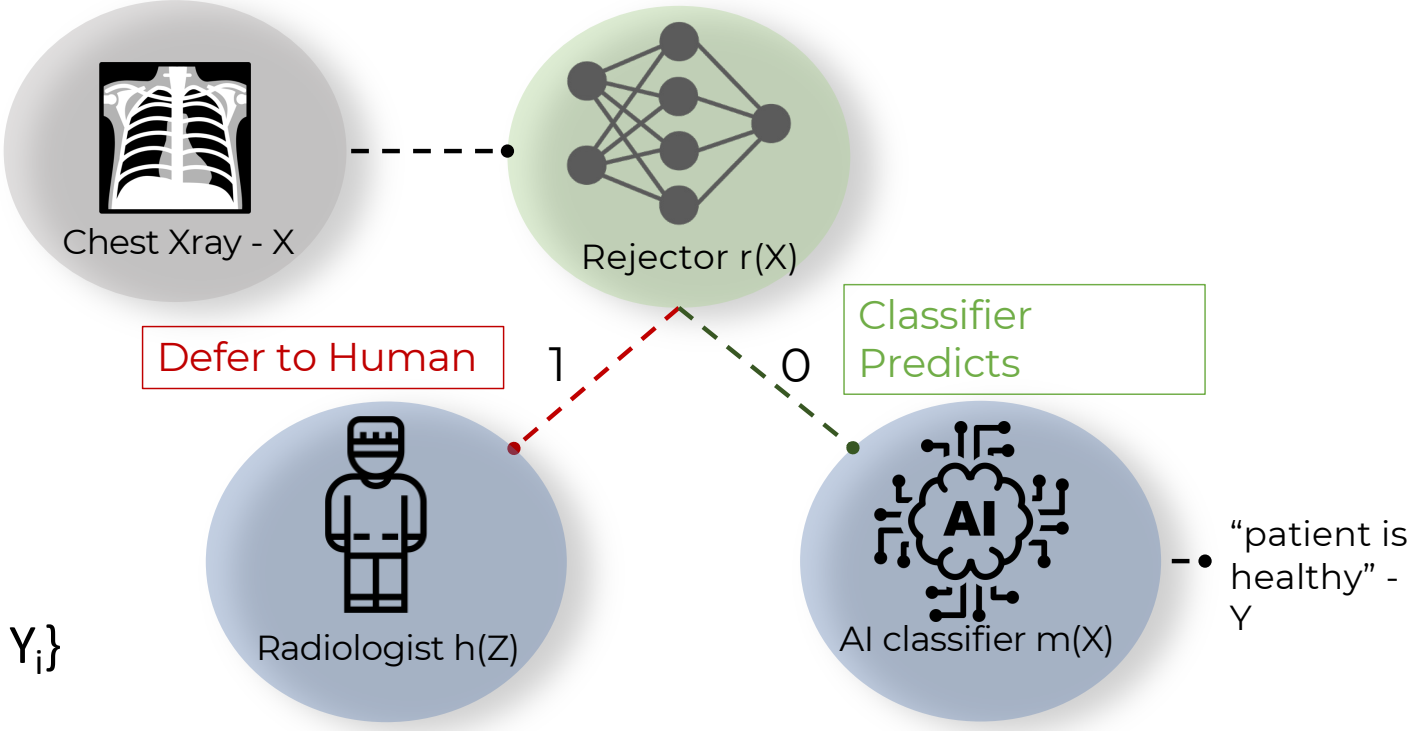
[1]:Beede, Emma, et al. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.

Takeaways from deployment

1. Protocols around use of model are crucial to its success
2. Human centered evaluation is crucial to be able to understand issues required for effective deployment

- Eliminating the ophthalmologists from the system removes safety checks against model failure (e.g., distribution shift) and input issues
- **Can do better by combining model and ophthalmologists then each alone!**

Deferral System



Data = $\{X_i, H_i, Y_i\}$

Objective: $$L_{\text{def}}^{0-1}(m, r) := \mathbb{P}\left[\left((1 - r(X))\, m(X) + r(X) h(Z)\right) \neq Y\right]$$

How to Learn a Deferral System

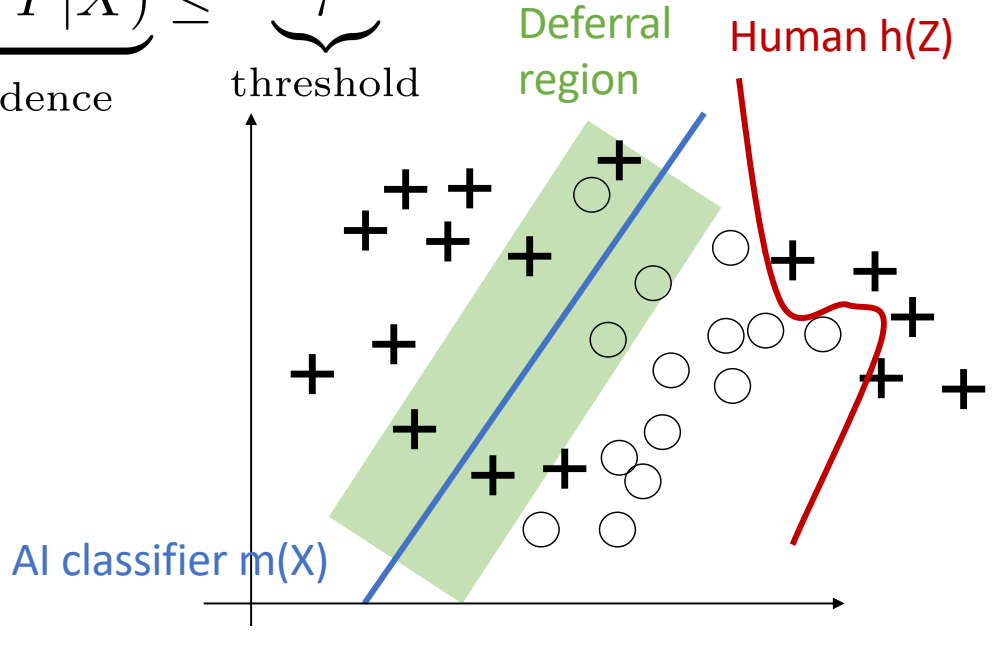
- **First Approach:** Threshold AI Classifier's confidence

$$\underbrace{r(X) = 1}_{\text{defer to human}} \iff \underbrace{\mathbb{P}(m(X) = Y|X)}_{\text{model confidence}} \leq \underbrace{\tau}_{\text{threshold}}$$

Doesn't consider humans' error!

-> Inside deferral region

- AI makes 3 mistakes
- **Human makes 5 mistakes!**



How to Learn a Deferral System

- **Better Approach:** Compare AI and Human confidence

$$\underbrace{r(X) = 1}_{\text{defer to human}} \iff \underbrace{\mathbb{P}(m(X) = Y|X)}_{\text{model confidence}} \leq \underbrace{\mathbb{P}(H = Y|X)}_{\text{estimated human confidence}}$$

How do we estimate **model confidence** and **human confidence**?

(Shown on blackboard.)

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, Sendhil Mullainathan. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort, 2019.

How to Learn a Deferral System

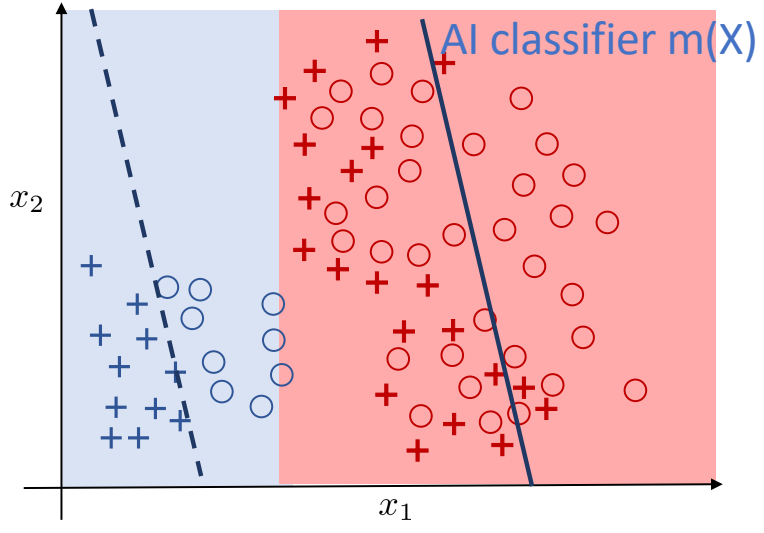
- **Better Approach:** Compare AI and Human confidence

$$\underbrace{r(X) = 1}_{\text{defer to human}} \iff \underbrace{\mathbb{P}(m(X) = Y|X)}_{\text{model confidence}} \leq \underbrace{\mathbb{P}(H = Y|X)}_{\text{estimated human confidence}}$$

Human perfect on red, bad at blue

Classifier does not adapt to Human!

- Classifier fit on average error tries to fit red group instead of red!



Jointly Learn Classifier and Rejector

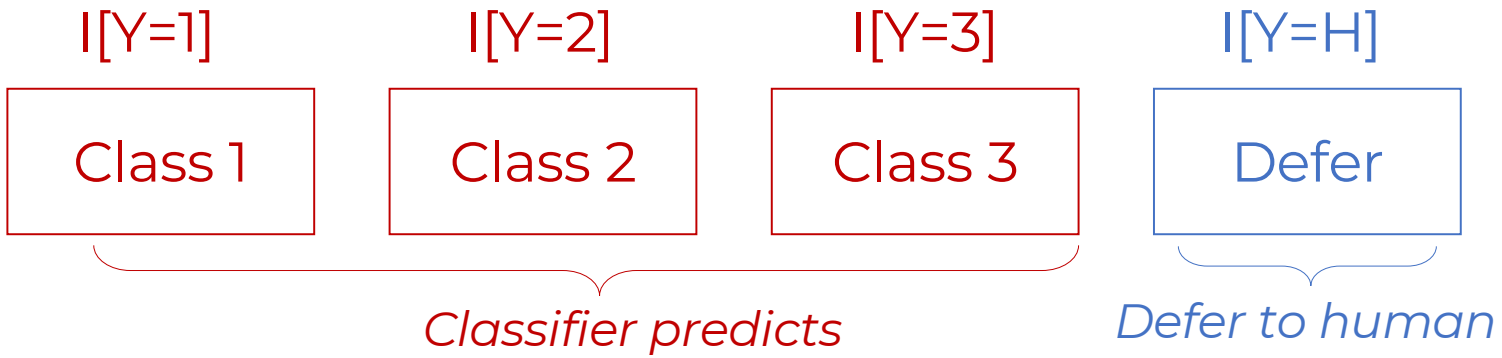
Learn Classifier and Rejector jointly!

- Optimize Classifier to **adapt** to Human's weaknesses and strengths
- Train Rejector to **defer** to who is more accurate between Human and classifier

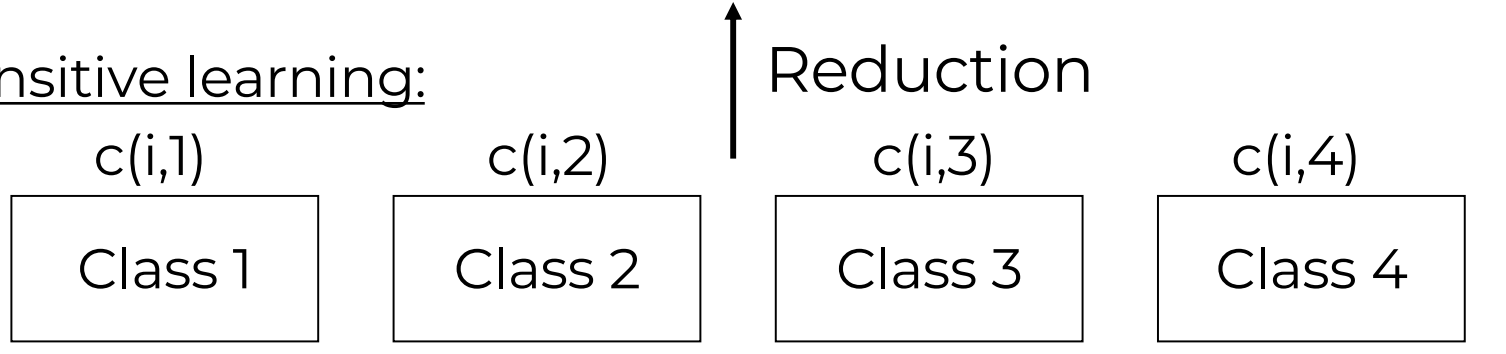
How to Practically Learn to Defer

At point i with X,H,Y:

Costs:



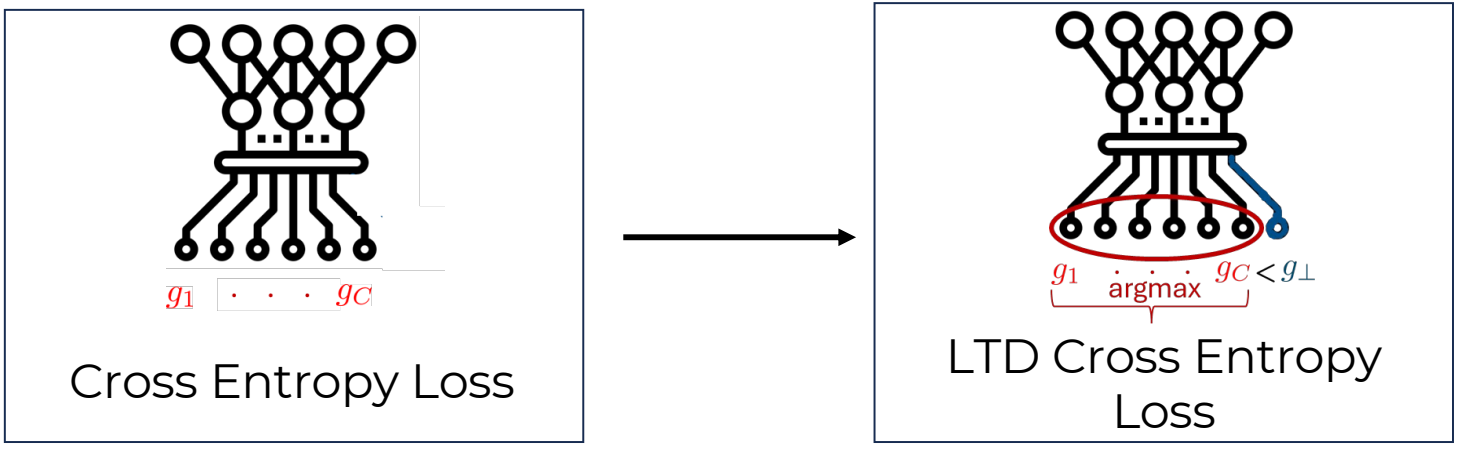
Cost sensitive learning:



Mozzanar, Sontag. Consistent Estimators for Learning to Defer to an Expert ICML 2020.

Cross Entropy Surrogate

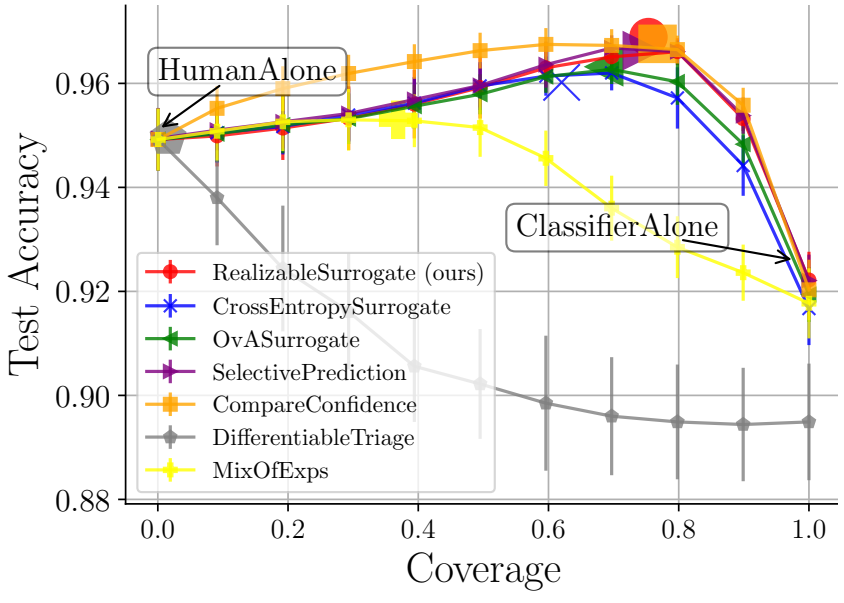
- **Theorem**: Any cost-sensitive classification loss can be adapted to learning to defer (LTD) and guarantee that it minimizes the LTD objective.



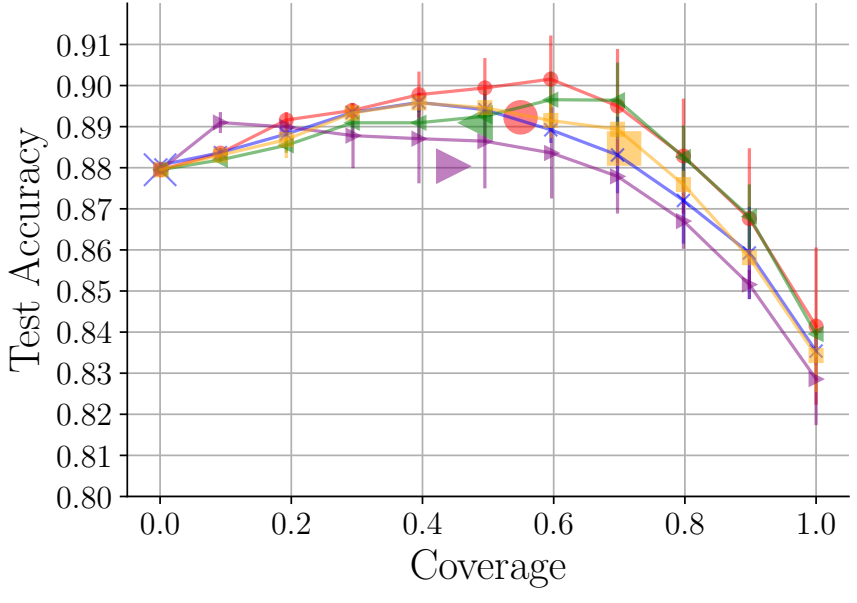
Standard ML

ML + Defer to Human

Mozzanar, Sontag. Consistent Estimators for Learning to Defer to an Expert ICML 2020.



(a) CIFAR-10H



(e) Chest X-ray - Airspace Opacity

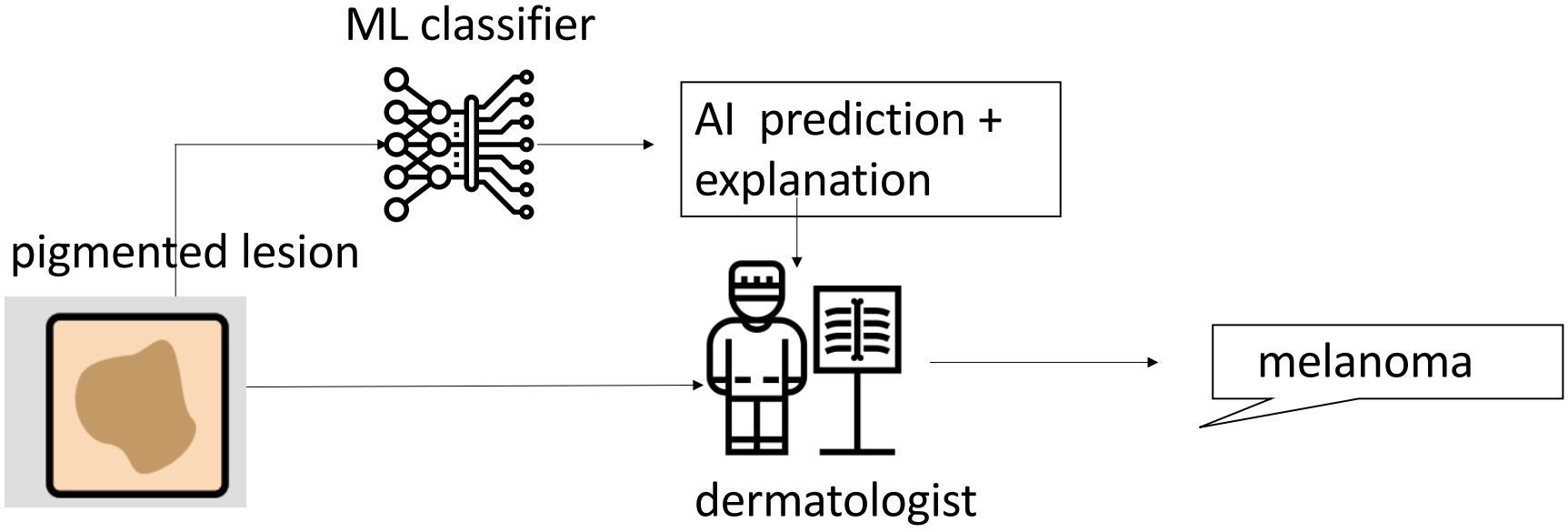
Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, David Sontag. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. AI & Stats 2023.

Triage can help towards automation

- The last iteration of the diabetic retinopathy project implemented this deferral setup with ungradable images being graded by an ophthalmologist.
- The human-AI team satisfies the constraints of the clinic, and if the rejector is chosen appropriately, can improve performance of the team
- **However, when clinician time is less scarce, we can allow for more explicit interaction between human-AI**

Model as a second opinion

Classify lesion into one of 7 categories: melanoma, …, vascular lesions [1]



[1]:Tschandl, Philipp, et al. "Human–computer collaboration for skin cancer recognition." *Nature Medicine* 26.8 (2020): 1229-1234.

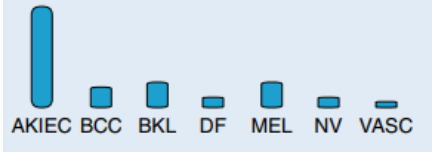
AI second opinion for skin cancer recognition

- 155 raters classified each 28 random images, and their performance (time and accuracy) was first measured (1) without AI and then (2) with AI predictions and explanations.

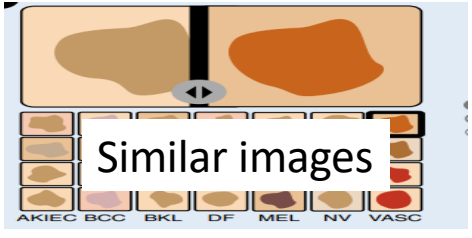
- Performance can vary based on two factors: 1) the AI explanations and 2) the specific dermatologist

[1]:Tschandl, Philipp, et al. "Human–computer collaboration for skin cancer recognition." *Nature Medicine* 26.8 (2020): 1229-1234.

Form of AI explanations has a big effect

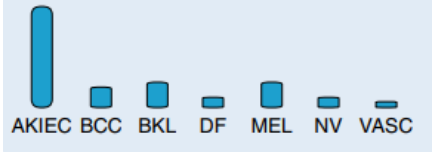


Multiclass probabilities

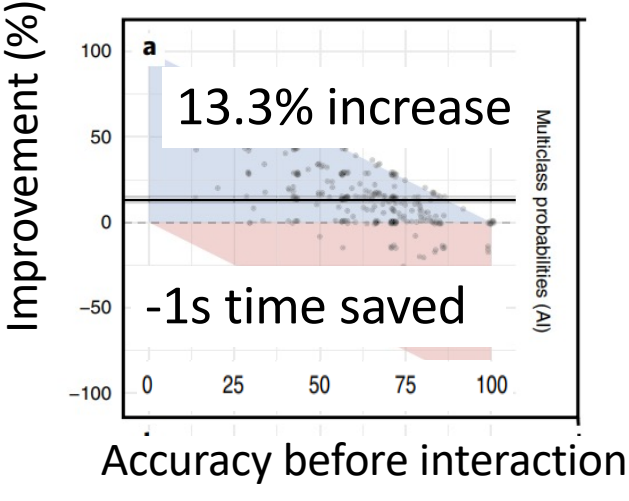


Which Explanation will clinicians benefit more from?

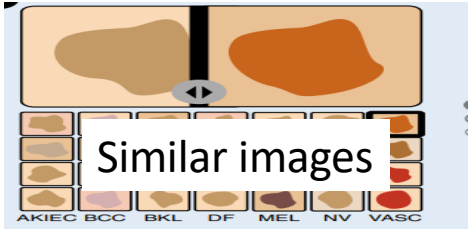
Form of AI explanations has a big effect

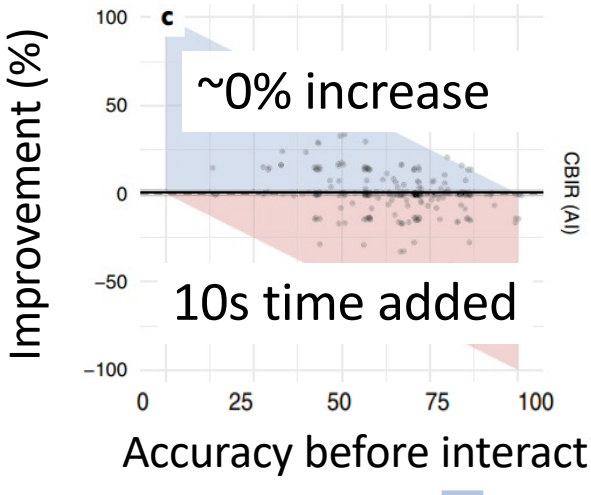


Multiclass probabilities



Accuracy before interaction



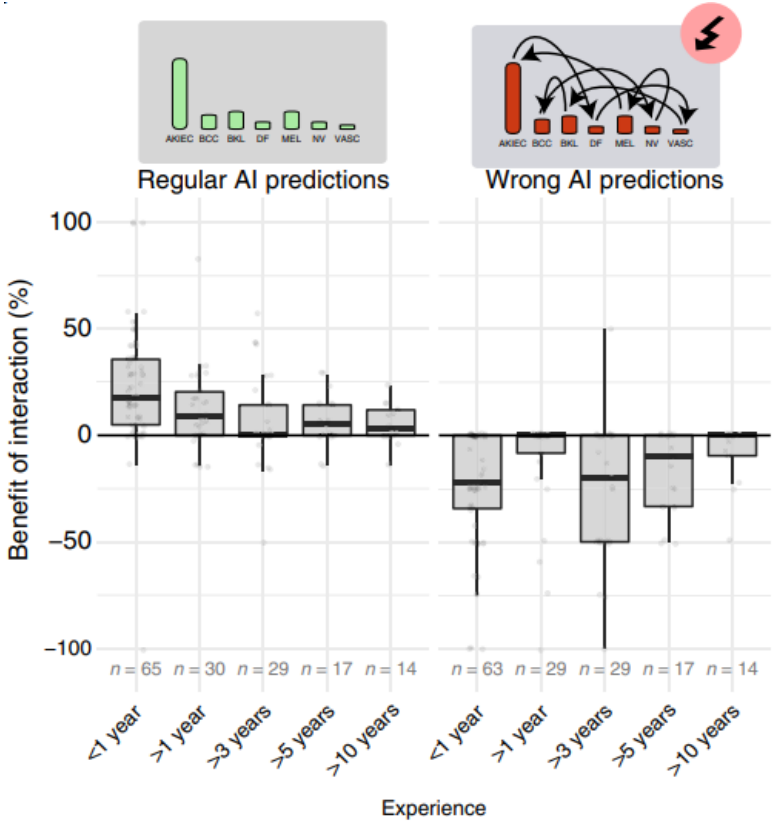


Accuracy before interaction

Possible benefit Possible loss

Clinician Experience and Confidence affects interactions

- Inexperienced raters benefit hugely from the regular AI, but are harmed the most from a bad AI model
- Experienced rater benefit the least from regular AI, and are harmed the Least by a bad AI model
- The difference is how sound their mental model of the AI is



Takeaways

Modes of human-AI interaction:

- Complete automation (AI only) or full human agency (no AI)
- Deferral System: AI delegates tasks to human or AI
- AI as a second opinion: AI gives the human a suggestion

Today's Lecture

1. How do we combine humans and AI?
   - Modes of human-AI interaction
2. **How do people think about AI?**
   - **Mental Models**
3. How do we interact with generative AI?
   - AI-assisted reading and writing

Mental Models

- **Mental model:** a person's understanding of how something works and how their actions affect it.
  - based on beliefs, flexible, limited and filters information.
  - sets expectation about what something can and cannot do and value can be gained from it
- What is special about **mental models of AI?**
  - Our priors are often wrong
  - AI's are evolving

Mental Model Formation

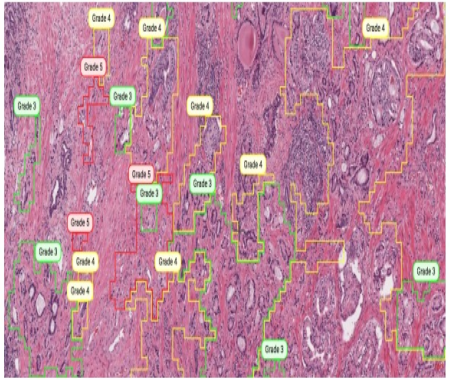
- How are mental models formed to begin with?

1. Through **experience**: as they interact with the AI more and more
2. Through **onboarding**: what we tell the human about the AI

What should we tell the human about the AI?

Study of Onboarding in Pathology

- 21 pathologists on task to understand prostate cancer risk [1]
- **Pre-Probe:** What types of information would you need to know about an AI assistant before using it?
- **Probe:** Diagnose a case with AI assistant
- **Post-probe:** What other information would you need to know about an AI assistant to work with it effectively?

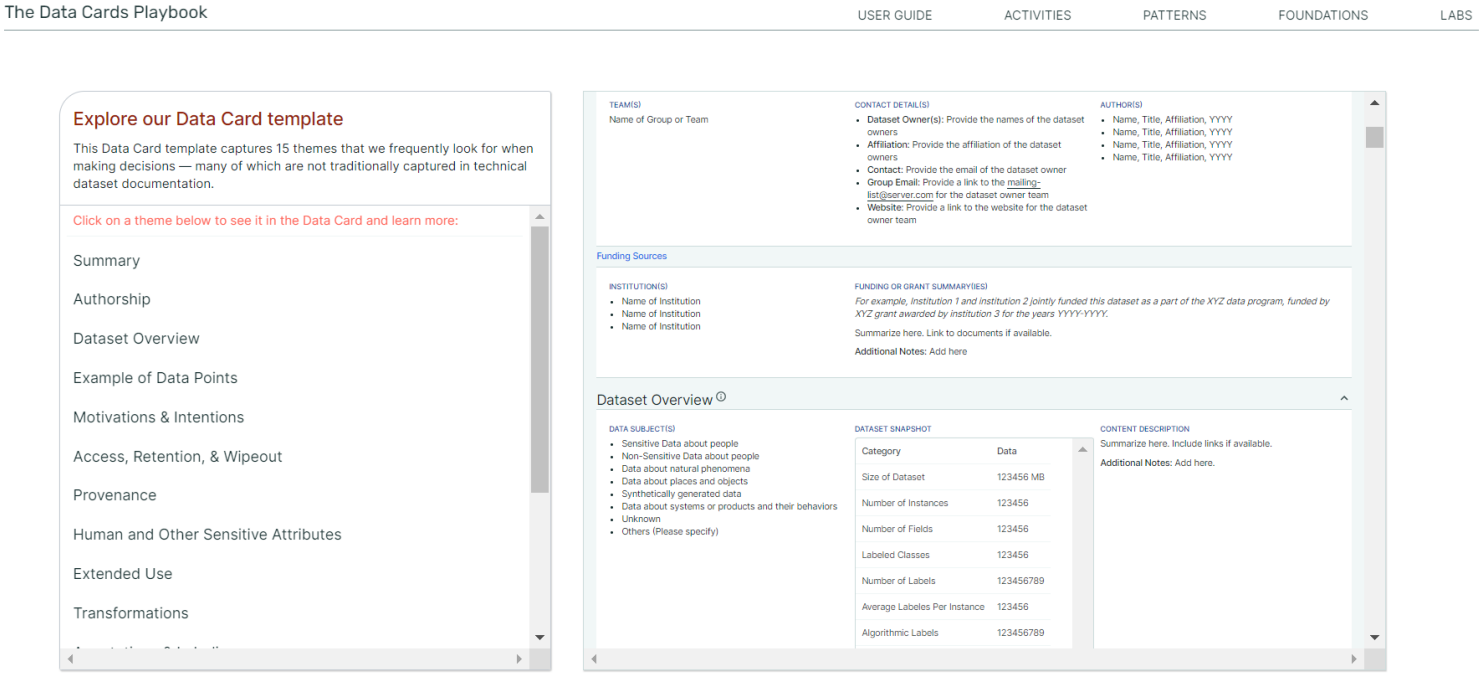


[1]:Cai, Carrie J., et al. "" Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making." Proceedings of the ACM on Human-computer Interaction 3.CSCW (2019): 1-24..

Training and Inference

- **Describe the scale of the training data.**
  - Some suggested that the number of data points should be on par with the volume of cases pathologists are typically trained on…
- **Describe the diversity of the training data.**
  - “More variation is better… Covering from community hospital to small groups, to academic medical centers”
- **Enumerate the data modalities that are accessible to the algorithm.**
  - “Does the AI assistant have access to information that I don’t have? Does it have access to any ancillary studies?”

Enable this with Data Cards



https://sites.research.google/datacardsplaybook/

Training and Inference

- **Specify the main steps of how the AI analyzes its inputs**
  - Some guessed it could only learn visual patterns derived from basic visual elements (“Maybe light and dark? Maybe colors? Maybe shapes, lines?”)
  - “Does it take into account the relationship between gland and stroma? Nuclear relationship?”
- **Specify where the algorithm received its source of ground truth.**
  - Participants asked whether the algorithm had learned from diagnoses made by general pathologists, GU pathologists, or an entire panel…

Calibration / “Point-of-View”

- **Demonstrate the subjective thresholds of the model using borderline cases.**
  - “I know what my friend… Will call… what would AI call it?… I’m treating it as a peer.”
- **Include a human-AI calibration phase.**
  - Pathologists envisioned assembling a set of cases with ground truth and comparing their diagnoses and the AI’s diagnoses with the ground truth in a calibration phase.

Accuracy and Performance

- **Define accuracy precisely.**
- **Provide human-relatable benchmarks for performance metrics**
  - Many were not sure what should constitute a reasonable performance threshold
- **Report AI performance on sub-categories of known human pitfalls**
  - “Maybe it has really good accuracy except for perineural invasion. If you see perineural invasion… Don’t fall for that.”

Enable this with Model Cards

Face Detection

Model Card v0 Cloud Vision API

- Overview
- Limitations
- Trade-offs
- Performance
- Test your own images
- Provide feedback

Explore

- Object Detection
- About Model Cards

Performance

Here you can dig into the model's performance on a selection of evaluation datasets drawn from different data sources than the training data. You can assess model performance across variables such as face size and facial orientation, as well as human-perceived skin tone, gender presentation, and age. Annotations for demographic variables were made by humans and used purely for testing; the model cannot detect them.

SUMMARY

- Area under the P-R curve (PR-AUC) is 0.84 (Open Images subset), 0.92 (Face Detection Dataset and Benchmark), and 0.94 (Labeled Faces in the Wild).
- Face size, facial orientation, and degree of occlusion all have a significant impact on model performance, with the model performing least well on faces that appear large (>25% of the image area), are looking to the left or right, and/or obstructed in some way.
- Disparities in recall are relatively small (< 3% gap) for all human-annotated demographic variables evaluated (perceived skin tone, gender presentation, age).

P-R CURVES

https://modelcards.withgoogle.com/face-detection and
https://huggingface.co/blog/model-cards

Can even describe in language the AI's ability compared to humans

- **Ideally**: create natural language rules (grounded in data) that describe how human should interact with AI

1. Rely on the AI for patients under the age of 40
2. Ignore on the AI whenever x-ray shows a pigtail catheters
3. Rely on AI whenever AI score is above 95%

....

Rules

Mozannar, Hussein, et al. "Effective Human-AI Teams via Learned Natural Language Rules and Onboarding." *Advances in Neural Information Processing Systems* 36 (2024).

What can happen if people have inaccurate mental models?

Today's Lecture

1. How do we combine humans and AI?
   - Modes of human-AI interaction
2. How do people think about AI?
   - Mental Models
3. **How do we interact with generative AI?**
   - **AI-assisted reading and writing**

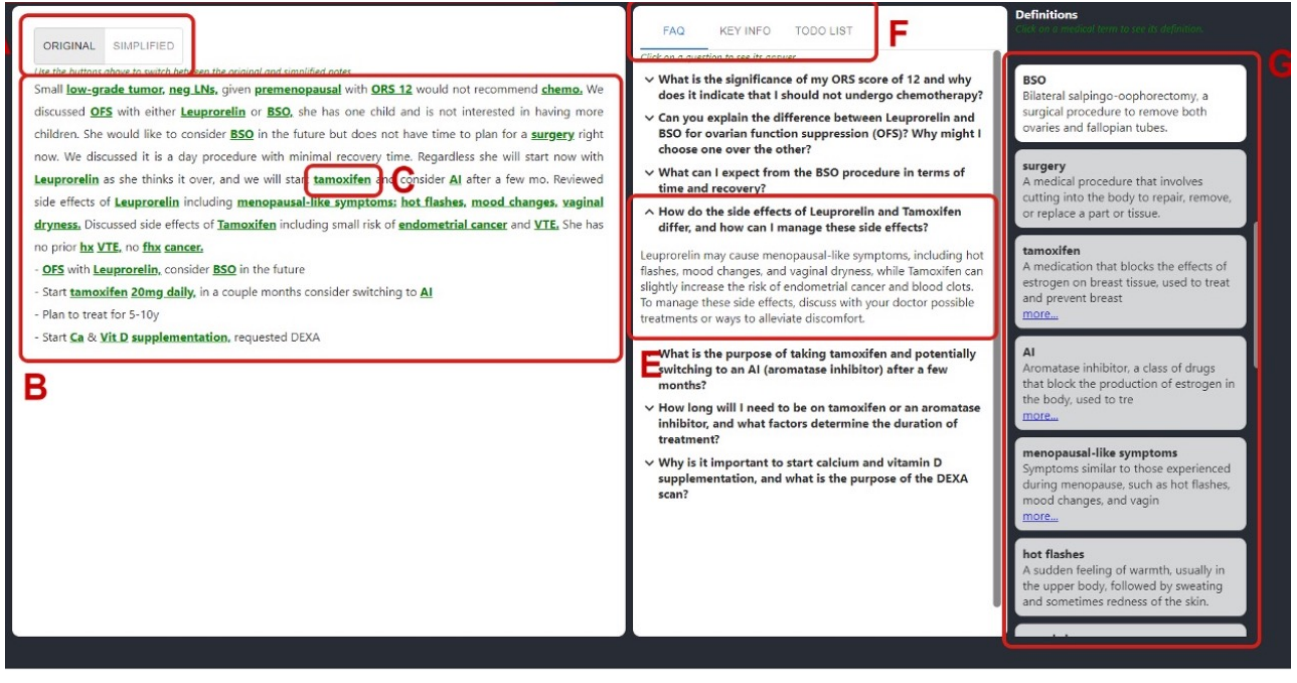
AI-Assisted Reading

- How we can use GPT-4 to help patients better understand their clinical notes?

42F with left breast DCIS high grade with comedo necrosis. Julia presents today for a second opinion. She has an appointment with Dr. Miller to discuss autologous reconstruction which is her preference. I agree with Dr. Anderson's advice and plans. I do think that she likely has extensive left breast DCIS as indicated by the several areas of suspicious calcifications and my assessment is that she will likely require mastectomy. I would like to review her case at our multidisciplinary tumor board including pathology review and evaluation of the contralateral breast or any areas of concern. I anticipate it will take us a week or so to get the slides from Cambridge so will tentatively plan on a presentation next week.

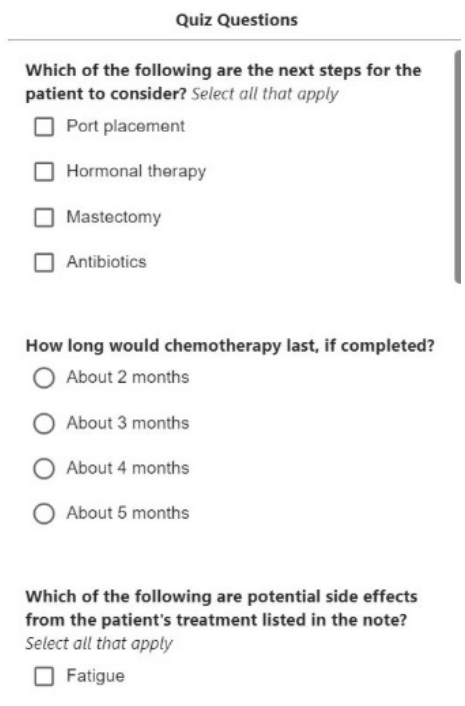
Mannhardt, N., Bondi-Kelly, E., Lam, B., O'Connell, C., Asiedu, M., Mozannar, H., ... & Sontag, D. (2024). Impact of Large Language Model Assistance on Patients Reading Clinical Notes: A Mixed-Methods Study. arXiv preprint arXiv:2401.09637.

Web interface powered by GPT-4 to help in reading clinical notes



Evaluation

- Given the note, participants (n=200) try to answer questions with and without the interface
- Participants were 20% more accurate when given the GPT-4 interface than without!
- However, GPT-4 definitions and answers often contained serious errors!

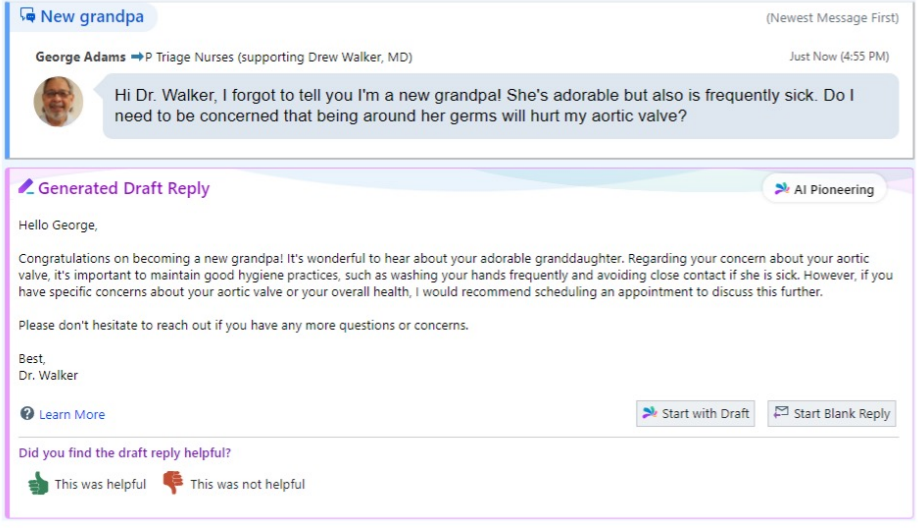


AI-Assisted Writing (from Lecture 7)

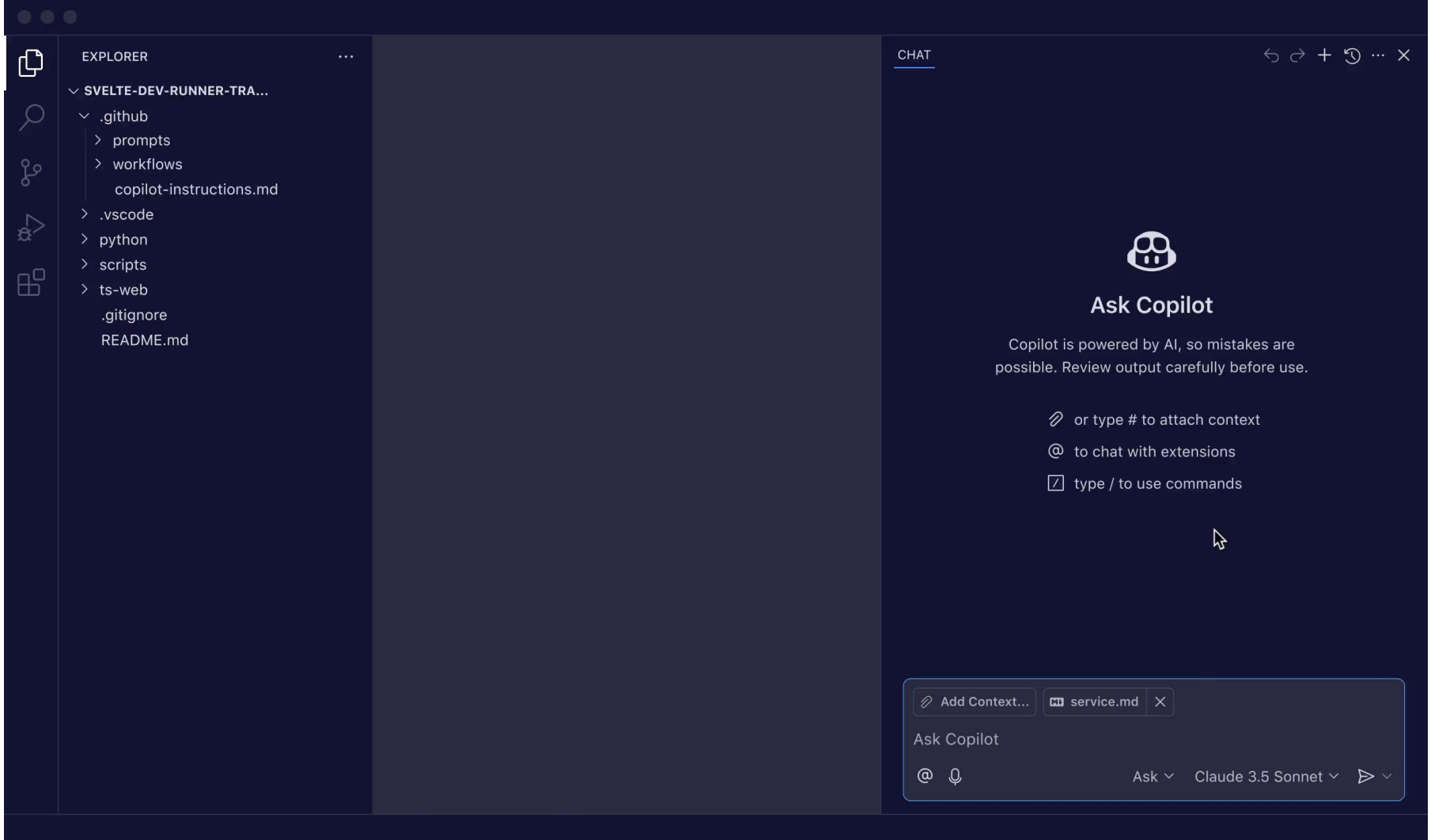
- How can LLMs help clinicians in writing clinical documentation or answering patient questions?

GPT-3.5/4 drafts messages to patients in the patient portal

- Clinicians used 20% of the drafts
- Food for thought – **what are other Clinician-AI interaction modalities?**



Garcia, Patricia, et al. "Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages." *JAMA Network Open* 7.3 (2024): e243201-e243201.


EXPLORER
SVELTE-DEV-RUNNER-TRA...
.github
prompts
workflows
copilot-instructions.md
.vscode
python
scripts
ts-web
.gitignore
README.md
CHAT
Ask Copilot
Copilot is powered by AI, so mistakes are possible. Review output carefully before use.
or type # to attach context
to chat with extensions
type / to use commands
Add Context...
service.md
Ask Copilot
Ask
Claude 3.5 Sonnet

Guidelines for Human AI Interaction

Learn more: https://aka.ms/aiguidelines

INITIALLY	1 Make clear what the system can do.	2 Make clear how well the system can do what it can do.					
DURING INTERACTION	3 Time services based on context.	4 Show contextually relevant information.	5 Match relevant social norms.	6 Mitigate social biases.			
WHEN WRONG	7 Support efficient invocation.	8 Support efficient dismissal.	9 Support efficient correction.	10 Scope services when in doubt.	11 Make clear why the system did what it did.		
OVER TIME	12 Remember recent interactions.	13 Learn from user behavior.	14 Update and adapt cautiously.	15 Encourage granular feedback.	16 Convey the consequences of user actions.	17 Provide global controls.	18 Notify users about changes.

Takeaways

- Figure out what mode of Human-AI interaction is appropriate for your problem
- Human's mental model of the AI determines the success of the system
- Design onboarding stages to allow the human to form an accurate mental model of the AI

Takeaways

- Design AI and AI explanations with human in mind to avoid over-reliance
- Allow for updates over time to interface and model to avoid under-reliance
- Integrate and evaluate LLMs to help patients/clinicians in their tasks