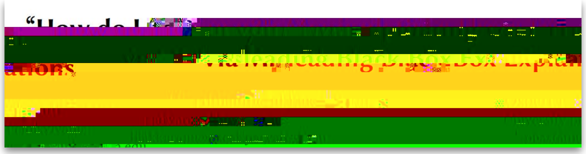
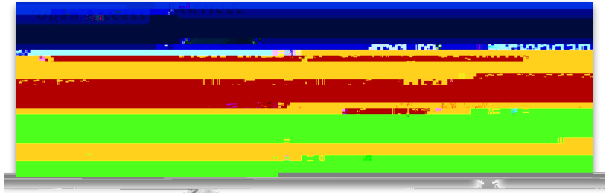
AI explanations can easily manipulate user’s trust [1]

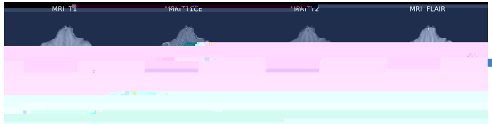


Explanations cannot help users detect potential model biases [2]

Explanations worsen physicians' task performance [3]



How can we make the AI explanations work as they are supposed to ?

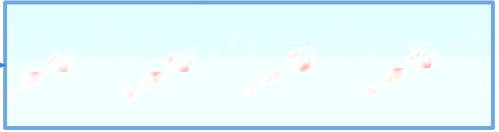


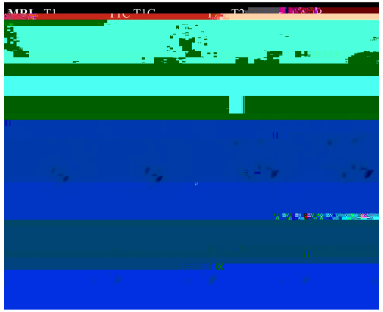
Input MRI

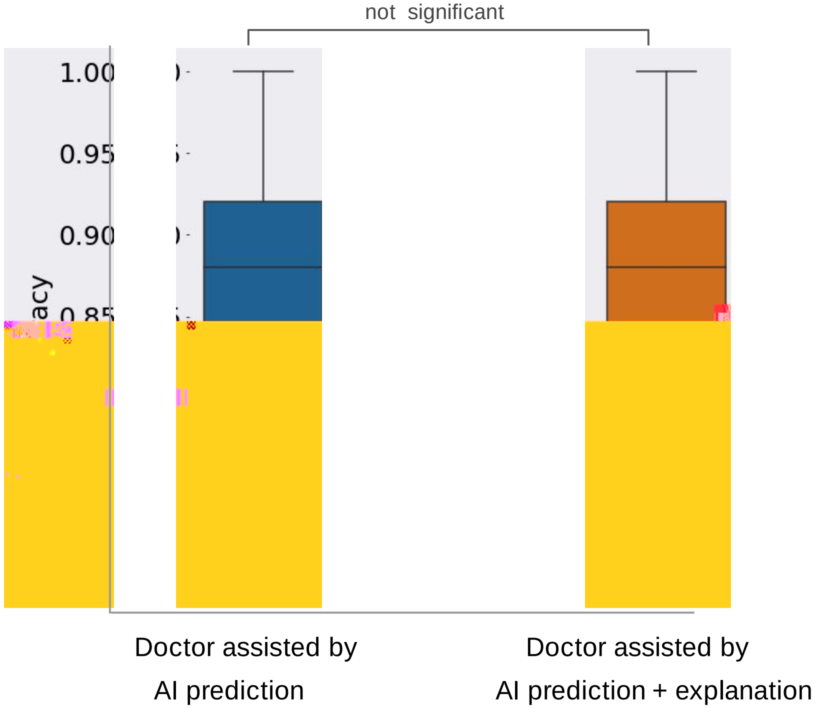
AI prototype

AI’s suggestion:
Grade 4 glioblastoma

AI’s explanation:







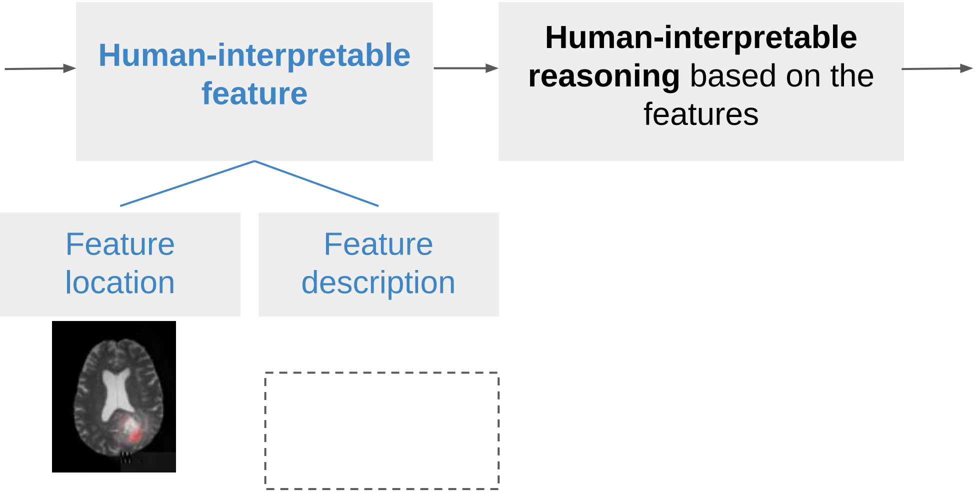
Quantitative results

AI explanations are

to improve doctors’ task performance

But why ?

> What (explanation) we get currently, when a radiologist read it, they **point out the significant features**, and then they **integrate those knowledge**, and say, to my best guess, this is a glioblastoma. And I have the same expectations of AI (explanation).
>
> – Neurosurgeon #3

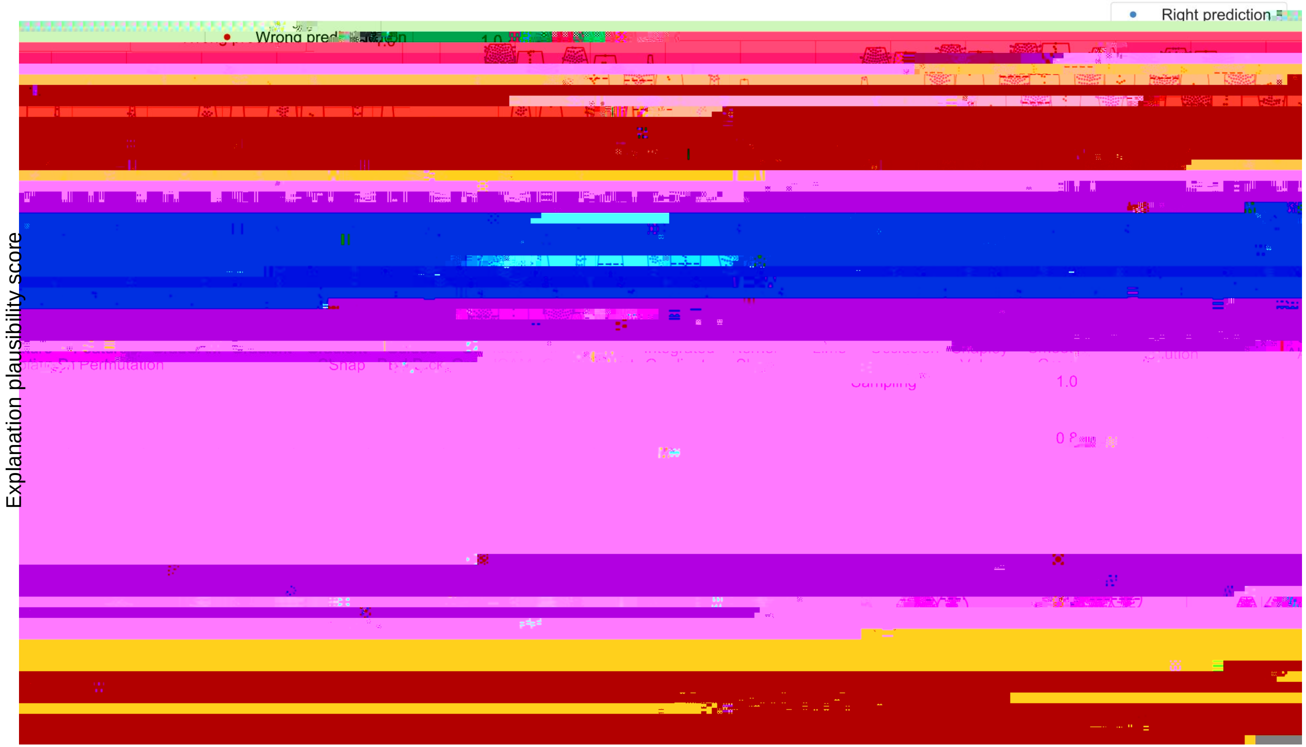


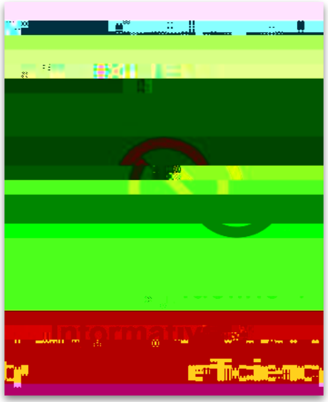
Human-interpretable feature

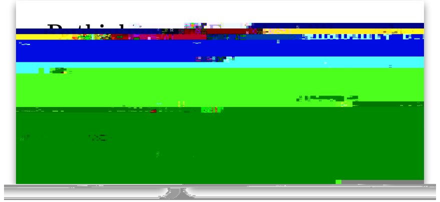
Human-interpretable reasoning based on the features

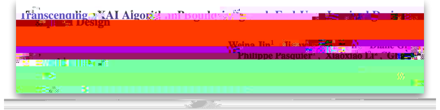
Feature location

Feature description


Right prediction
Wrong prediction
Explanation plausibility score
Permutation
Shap
Integrated
Lime
Occlusion
Sampling
1.0
0.8


Information
efficiency

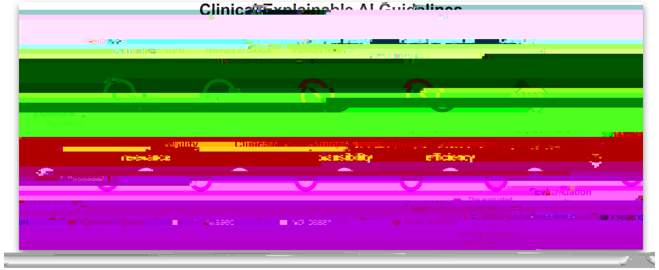




Transcending XAI Algorithm Boundaries through End-User-Inspired Design

Weina Jin[1], ... Diane Gromala ...

Philippe Pasquier[1], Xiaoxiao Li[2], Gh...



Clinical Explainable AI Guidelines

End-User-Centered Explainable AI Framework

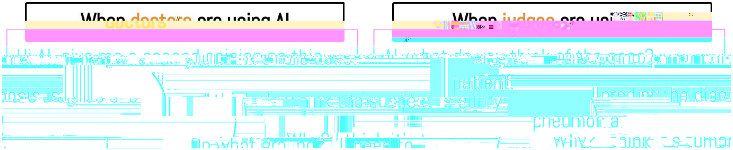
Thank you!

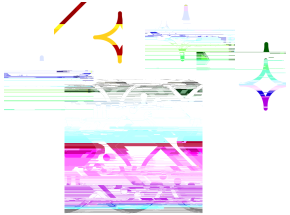
How technologies are ignoring values from underrepresented groups and how we combat it

[illegible]

XAI ignores end-users by:	**What is it?**	**Why is it harmful?**	**How we combat it?**
1. Not aligning with **human reasoning** and interpretation patterns with explanation	Explanations have incomplete feature description only feature localization or text description, not both	Users can hardly incorporate evidence from explanations into their decision process	Design new XAI techniques to provide explanation with complete feature description [Work in progress]
2. Not following **human communication norms** with explanations	Explanations are created to be plausible *regardless of AI decision correctness*	Users in critical tasks can have worse performance that harms people’s life, money, etc.	Reveal to the XAI community such ill practice and its harmfulness [1]
3. Not being designed to fulfill users’ **utility of explanation**	XAI algorithms are not designed for its utility to end-users, e.g., verifying AI decisions, ensuring AI safety, and improving human-AI performance	Cannot effectively help uses to solve their problems when seeking explanations	Propose user-centered XAI evaluation objectives and metrics [2,3]

Explainability needs to be carefully crafted based on end-user-centered requirements.


When doctors are using AI
When judges are using AI
Hi AI, give me a second opinion on this
AI, what do you think of the case?
patient
I predict the diagnosis is
I think the suspect is guilty.
pneumonia.
On what ground? I need to
Why? I think it's

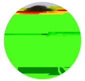


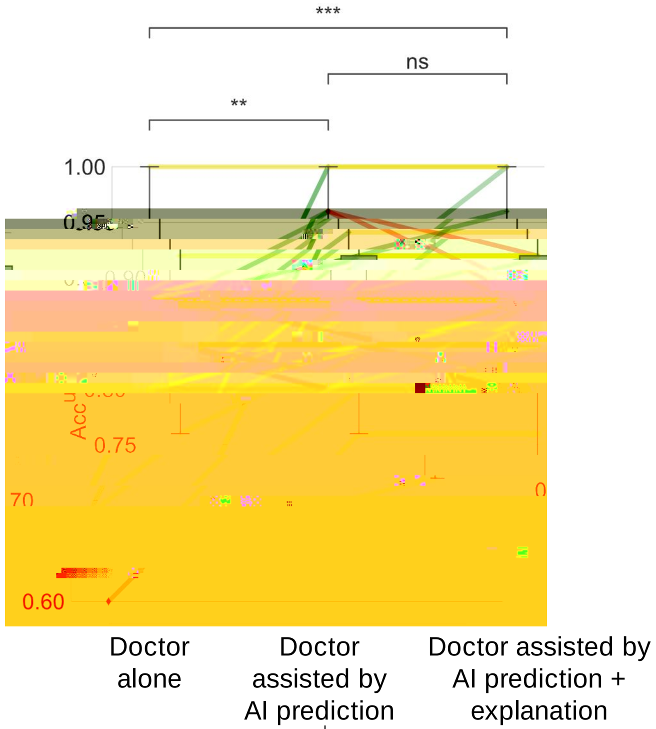
Explainable AI

- Trustworthiness
- Accountability
- Improving task performance

How technologies are ignoring values from underrepresented groups and how we combat it

Team:

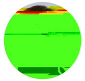



***
ns
**
1.00
0.95
Acc
0.75
70
0.60
Doctor alone
Doctor assisted by AI prediction
Doctor assisted by AI prediction + explanation

How technologies are ignoring values from underrepresented groups and how we combat it

Team:

Advisors:



We want AI to be safe, reliable, and accountable to use

HAMAMATSU

Medical Image

SIMON FRASER

> “
>
> Algorithms are designed by people, and
>
> . It’s rarely intentional—but this doesn’t mean we should let data scientists off the hook. It means we should be critical about and vigilant for the things we know can go wrong. If we assume discrimination is the default, then we can design systems that work toward notions of equality. [1]