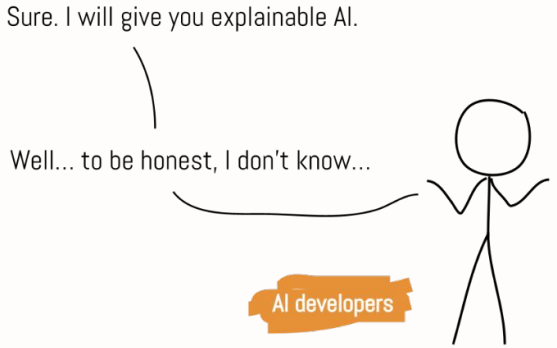
Clinical User-Centred Explainable AI for Medical Image Analysis

A walkthrough of an end user-centered design practice





TA: Weina Jin, MD

PhD Candidate
Medical Imaging Analysis Lab
School of Computing Science
Simon Fraser University

HAMARNEH
Medical Image Analysis Research Group

SFU SCHOOL OF COMPUTING SCIENCE

Core questions

1. What are end user-centered requirements for explainable AI algorithms?
2. How to design and evaluate clinical user-centered explainable AI?

My user-centered approach

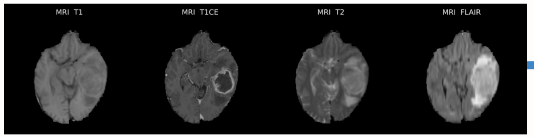
1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors



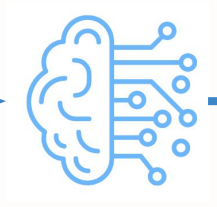
Medical image analysis

Use computational methods to interpret and analyze information from medical images

For example, in clinical decision support:



Input
brain image



Model

Suggestion from the model:
High-grade brain tumor

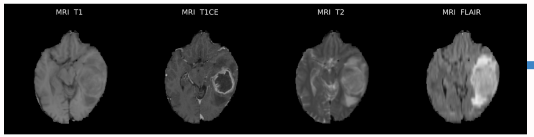
Output



Medical image analysis

Use computational methods to interpret and analyze information from medical images

For example, in clinical decision support:



Input brain image

Model

Suggestion from the model:
High-grade brain tumor

Output



AI models are largely black-box, human-uninterpretable models because of

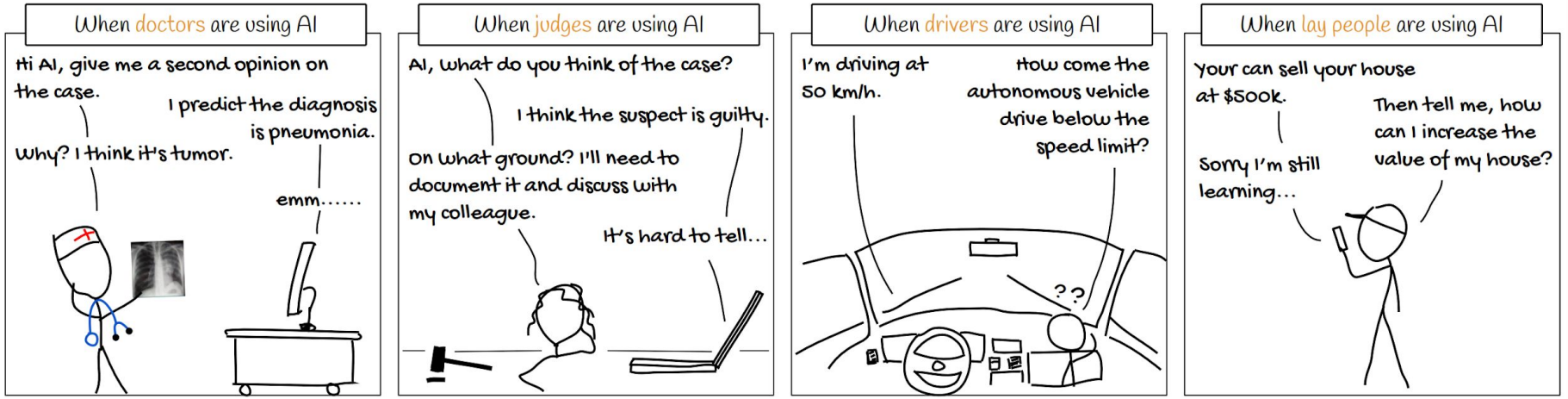
- Millions of parameters
- Many non-linear transformations from input to output

Explainable AI (XAI)

In critical tasks that are related to humans' life, money, career, etc.,

for humans to make informed decision assisted by a model,

the model needs to explain the predictions in human-understandable ways [1].



[1] Towards A Rigorous Science of Interpretable Machine Learning

Potential clinical utilities of AI explanations

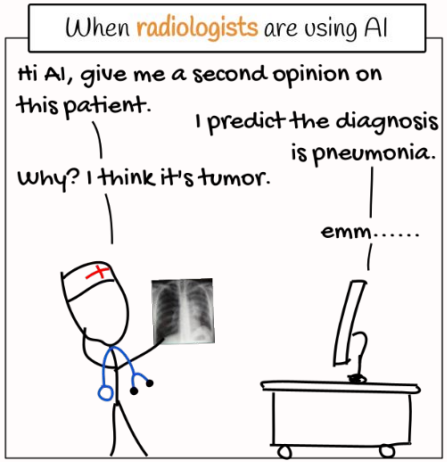
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

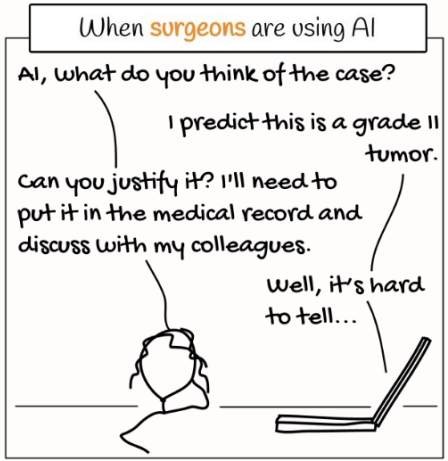
Resolve clinical users' disagreement with AI [1]
informed decision

Verify AI's suggestion [1]
evidence-based medicine

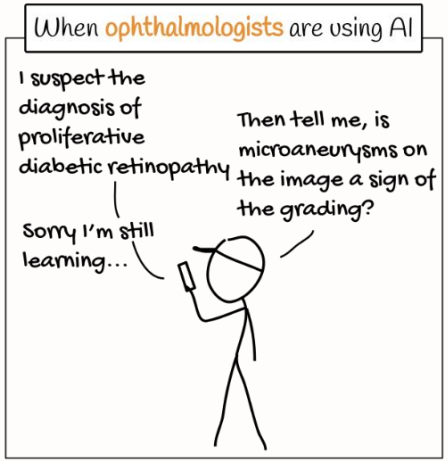
Calibrate clinical users' trust [1]

Learning & medical knowledge discovery [1]









[1] EUCA: the End-User-Centered Explainable AI Prototyping Framework. arXiv:2102.02437
[2] Towards A Rigorous Science of Interpretable Machine Learning

What does "clinical user-centered" really mean?

XAI is inherently a **human-centered** concept.

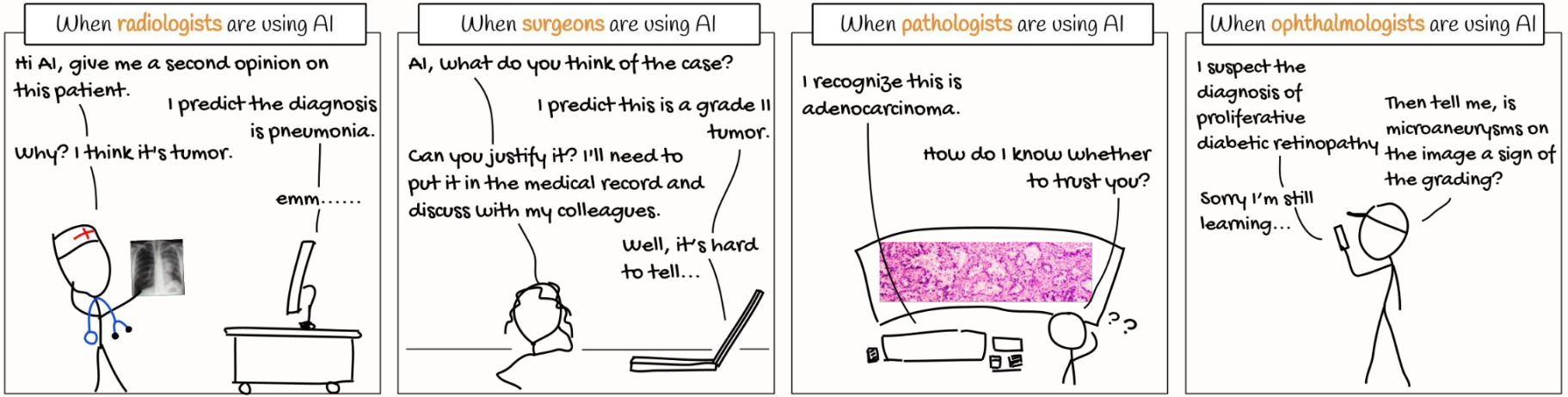
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

Resolve clinical users' disagreement with AI [1]
informed decision

Verify AI's suggestion [1]
evidence-based medicine

Calibrate clinical users' trust [1]

Learning & medical knowledge discovery [1]



[1] EUCA: the End-User-Centered Explainable AI Prototyping Framework. arXiv:2102.02437
[2] Towards A Rigorous Science of Interpretable Machine Learning

Core questions

1. What are end user-centered requirements for explainable AI (XAI) algorithms?
2. How to design and evaluate clinical user-centered explainable AI (XAI)?

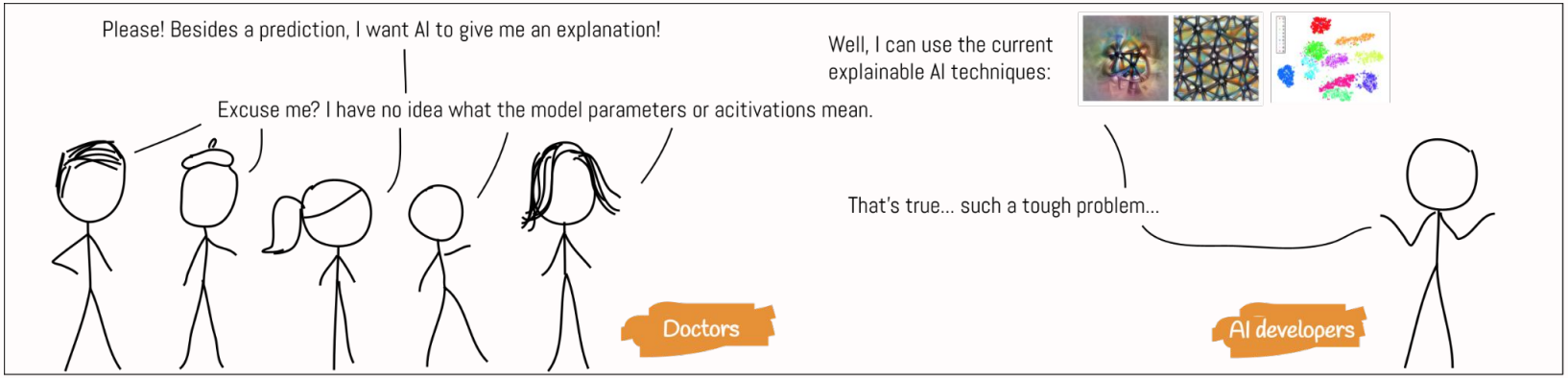
My user-centered approach

1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors



- Usability evaluation

What does “clinical user-centered ” really mean?

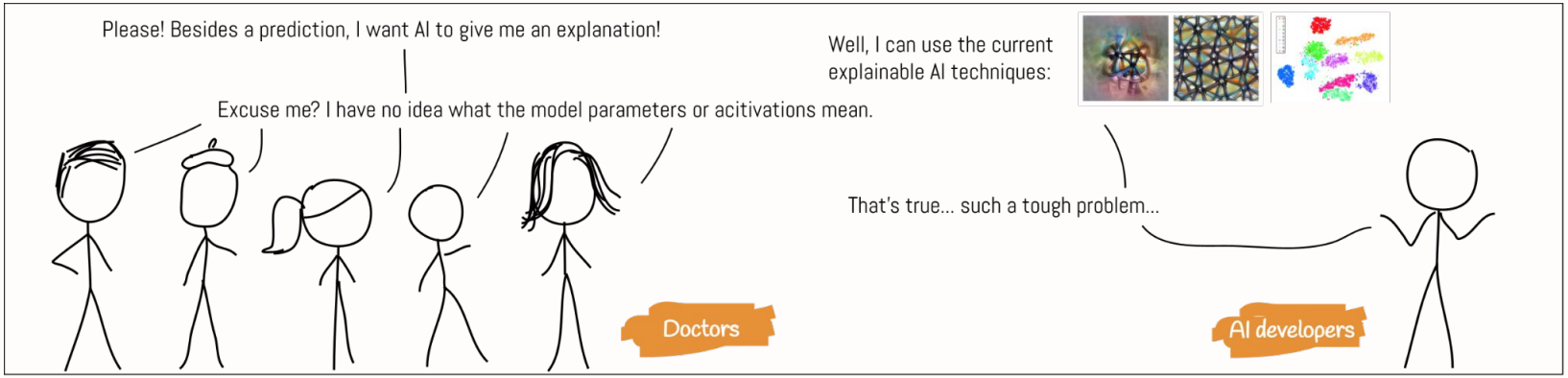


Some XAI techniques are not end-user friendly:

Prerequisite technical knowledge in AI/ML/data science

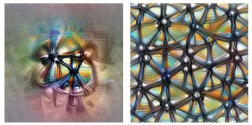
- Usability evaluation

What does “clinical user-centered ” really mean?



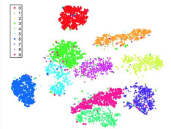
Some XAI techniques are not end-user friendly:

Prerequisite technical knowledge in AI/ML/data science



Prerequisite technical knowledge: understand the role of neurons in model architecture

Neuron Visualization.
Olah et al. Distill. 2017



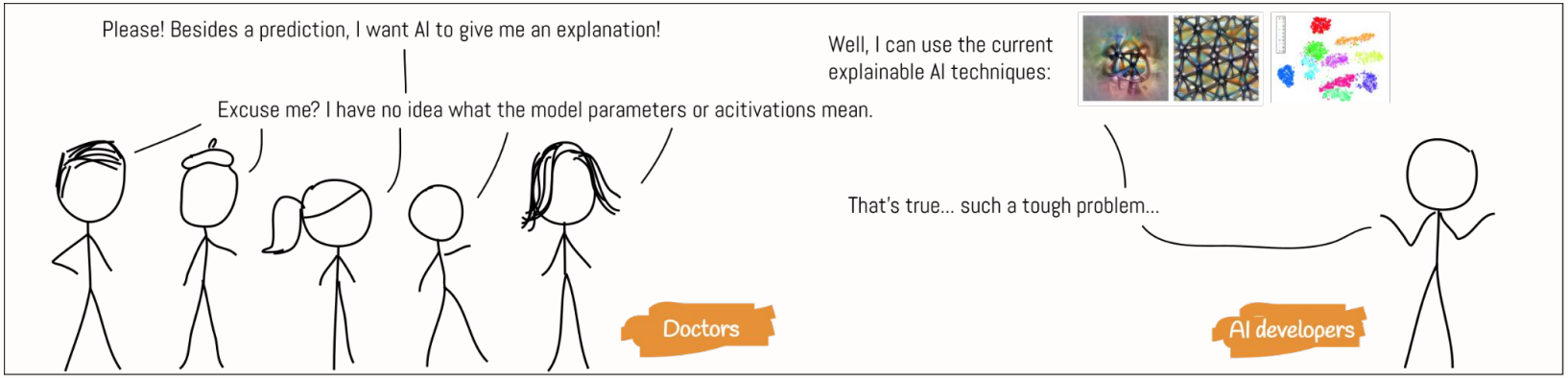
Prerequisite technical knowledge: understand assumptions of a clustering method

Visualizing Data using t-SNE.
Laurens van der Maaten and Geoffrey Hinton. JMLR. 2008

- **Usability evaluation**

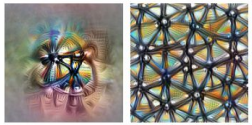
What does “clinical user-centered ” really mean?

The explanation form is understandable without technical knowledge



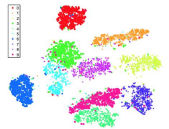
Some XAI techniques are not end-user friendly:

Prerequisite technical knowledge in AI/ML/data science



Prerequisite technical knowledge: understand the role of neurons in model architecture

Neuron Visualization. Olah et al. Distill. 2017



Prerequisite technical knowledge: understand assumptions of a clustering method

Visualizing Data using t-SNE. Laurens van der Maaten and Geoffrey Hinton. JMLR. 2008

No technical knowledge is required to understand the explanation

Guideline 1

Understandability

- **Usability evaluation output**

What does “clinical user-centered ” really mean?

Various explanation forms
(and their corresponding XAI algorithms)

The explanation form is understandable without technical knowledge



End-user-friendly explanation forms

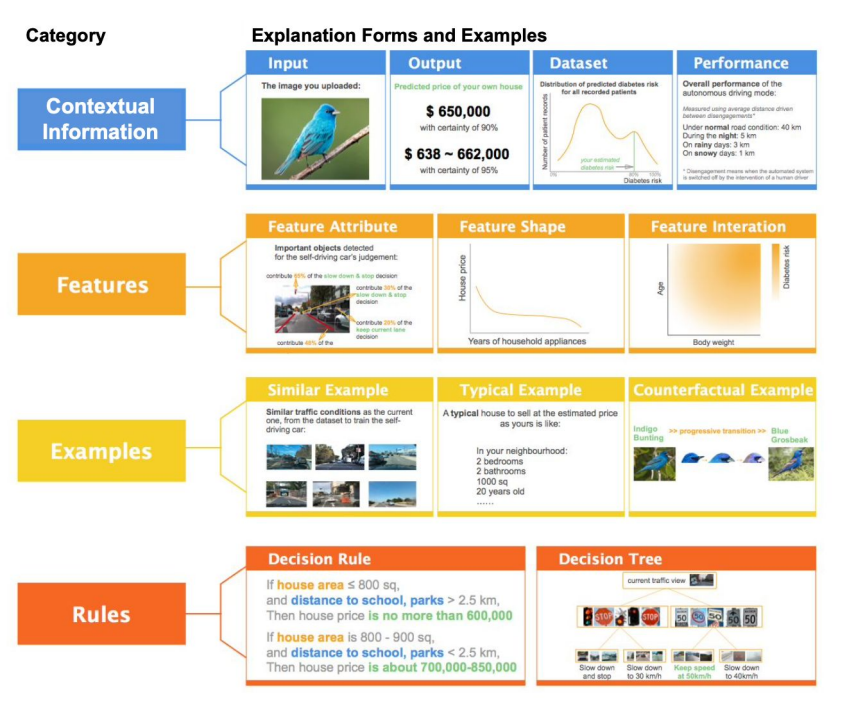
● Usability evaluation result

What does “clinical user-centered ” really mean?

Various explanation forms
(and their corresponding XAI algorithms)

The explanation form is understandable without technical knowledge

End-user-friendly explanation forms



EUCA: the End-User-Centered Explainable AI Prototyping Framework. arXiv:2102.02437

● Usability evaluation result

EUCA: the End-User-Centered Explainable AI Framework



User requirements

Backend XAI algorithms

Explanation Form	End Users' Interpretation & Requirement	Visual Representation	XAI Algorithm
Feature attribution	Can answer how and why AI reaches its decisions; Can assess AI by comparing features with user's prior knowledge; Users may require different levels of detail on the attribution	Saliency map; Bar chart	Linear or scoring models (Ustun and Rudin 2016), Shapley value (Shapley 1951), SHAP (Lundberg and Lee 2017), TCAV (Kim et al. 2018), LIME (Ribeiro, Singh, and Guestrin 2016), CAM (Zhou et al. 2015)
Feature shape	Graphic representation is easy to understand; Support counterfactual reasoning	Line plot	GAM (Tan et al. 2023), PDP (Friedman 2001), ALE (Apley 2016), SHAP dependence (Lundberg, Allen, and Lee 2017)
Feature interaction	Difficult to interpret; Users require to prioritize significant feature interactions	2D or 3D heatmap	GA2M (Caruana et al. 2015), PDP (Friedman 2001), ALE (Apley 2016), SHAP interaction (Lundberg, Allen, and Lee 2017)
Similar example	Similar to human decision-making; Users require support for side-by-side feature-based comparison among examples	Data instance	*k*-nearest neighbors, graph-based approach (Kawahara, Moriarty, and Hamarneh 2017), CBR (Kolodner 1992)
Typical example	May be misinterpreted as similar example; Users require atypical examples to show edge cases for safety and biases check	Data instance	*k*-mediods, MMD-critic (Kim, Khanna, and Koyejo 2016), CNN prototype (Li et al. 2018; Chen et al. 2019; Simonyan, Vedaldi, and Zisserman 2013; Mahendran and Vedaldi 2014)
Counterfactual example	Can suggest improvements, and help users differentiate similar instances; Users require to be able to define the counterfactual outcome, the counterfactual feature type and range	Data instance	Visual counterfactual (Goyal et al. 2019), pertinent negatives (Dhurandhar et al. 2018)
Decision rule	Decision logic is simple and understandable; Users require XAI to balance explanation completeness and simplicity	Text, table, or matrix	Bayesian rule lists (Yang, Rudin, and Seltzer 2017), LORE (Guidotti et al. 2018), Anchors (Ribeiro, Singh, and Guestrin 2018)
Decision tree	Difficult to interpret; Can support counterfactual reasoning; Users require XAI to highlight branches for users' interested instances	Tree diagram	Disentangled CNNs (Zhang et al. 2018); model distillation (Frosst and Hinton 2017)

EUCA: the End-User-Centered Explainable AI Prototyping Framework. arXiv:2102.02437

Core questions

1. What are end user-centered requirements for explainable AI (XAI) algorithms?
2. How to design and evaluate clinical user-centered explainable AI (XAI)?

My user-centered approach

1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors



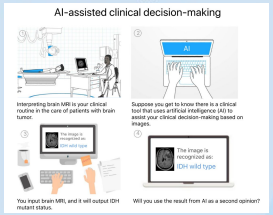
- Card-based low-fidelity prototype

Co-design XAI with doctors: study procedure

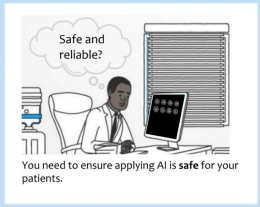
Round 1 (RQ2) End Users' Requirements for Explanation Goals

Round 2 (RQ1) End Users' Requirements for Explanation Forms

Present the AI-assisted decision-making task

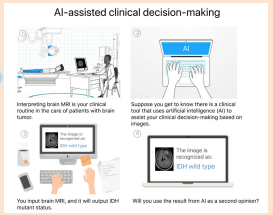


Present the explanation goal

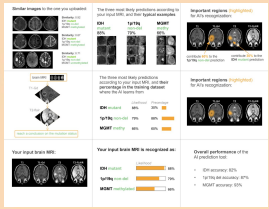


For each explanation goal, ask the participant whether & what kind of explanation/information is needed

Present the AI-assisted decision-making task again



Walk through the cards of explanation forms

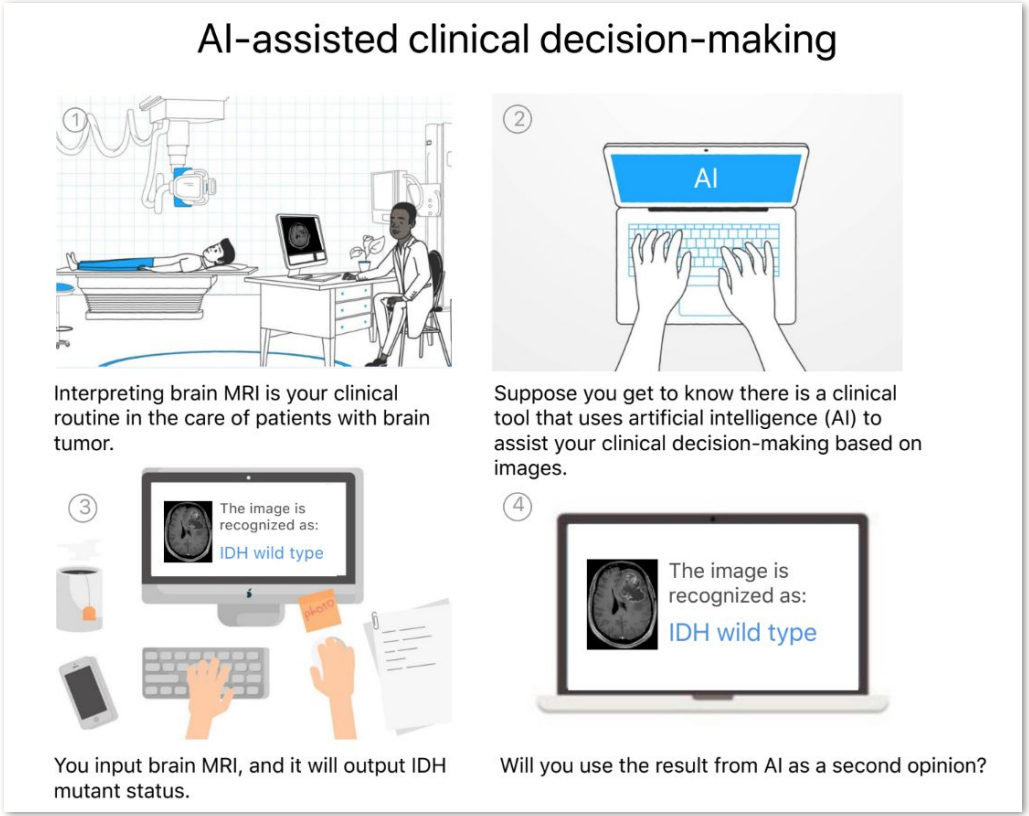


For each explanation goal, select, rank, & comment the explanation forms



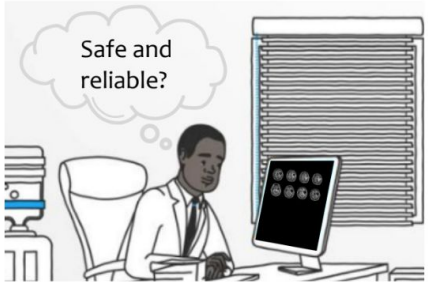
- Card-based low-fidelity prototype

Co-design XAI with doctors: the AI-assisted decision-making task



- Card-based low-fidelity prototype

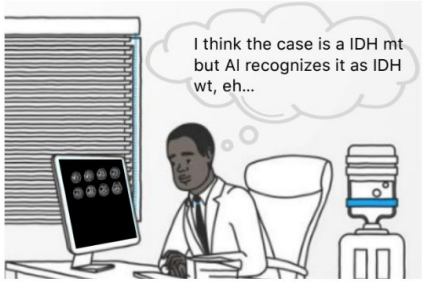
Co-design XAI with doctors: explanation goals



You need to ensure applying AI is **safe** for your patients.



You want to ensure the decision is made to the best interest of the patient and the patient will present high adherence.



The results sometimes do not align with your expectation.



In differential diagnosis, you want to make sure ***all possible conditions*** are included.



You doubt whether the AI tool will perform equally the same among people with ***different gender, age, or ethnicity group***.



You want to know which factors are ***justifiable*** to improve the patient's outcome.

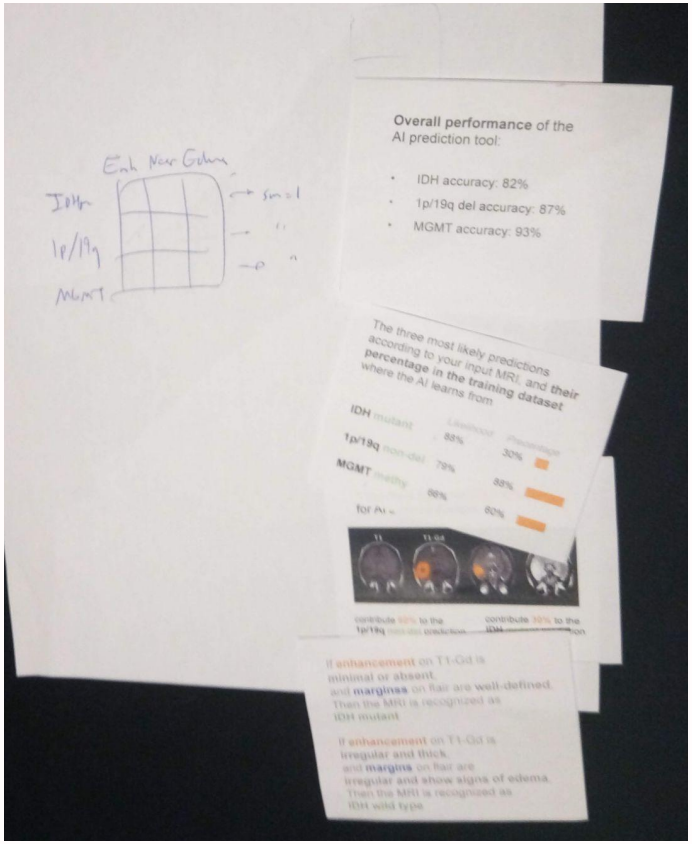
- Card-based low-fidelity prototype

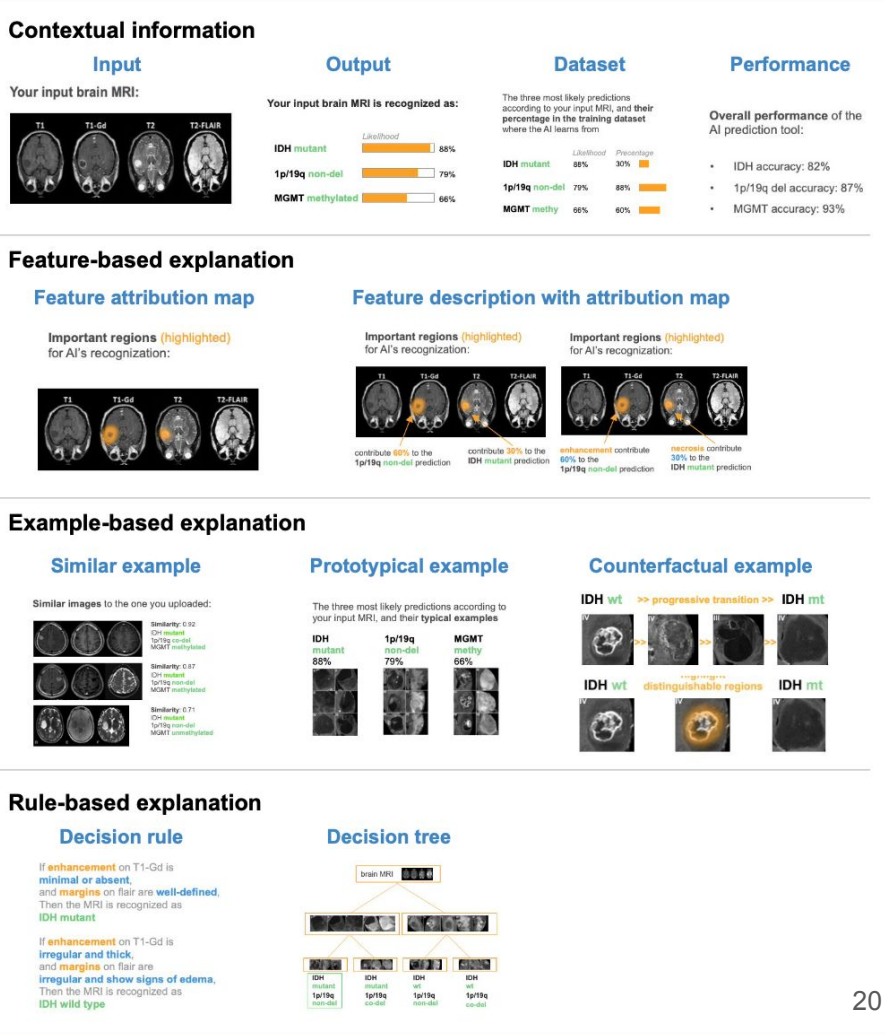
Co-design XAI with doctors: explanation forms



- Card-based low-fidelity prototype

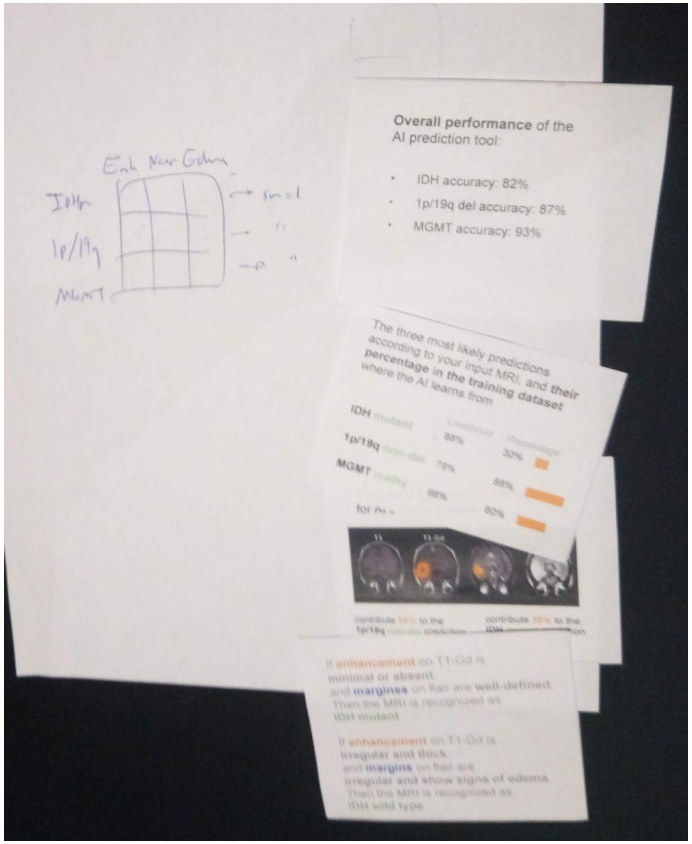
Co-design XAI with doctors

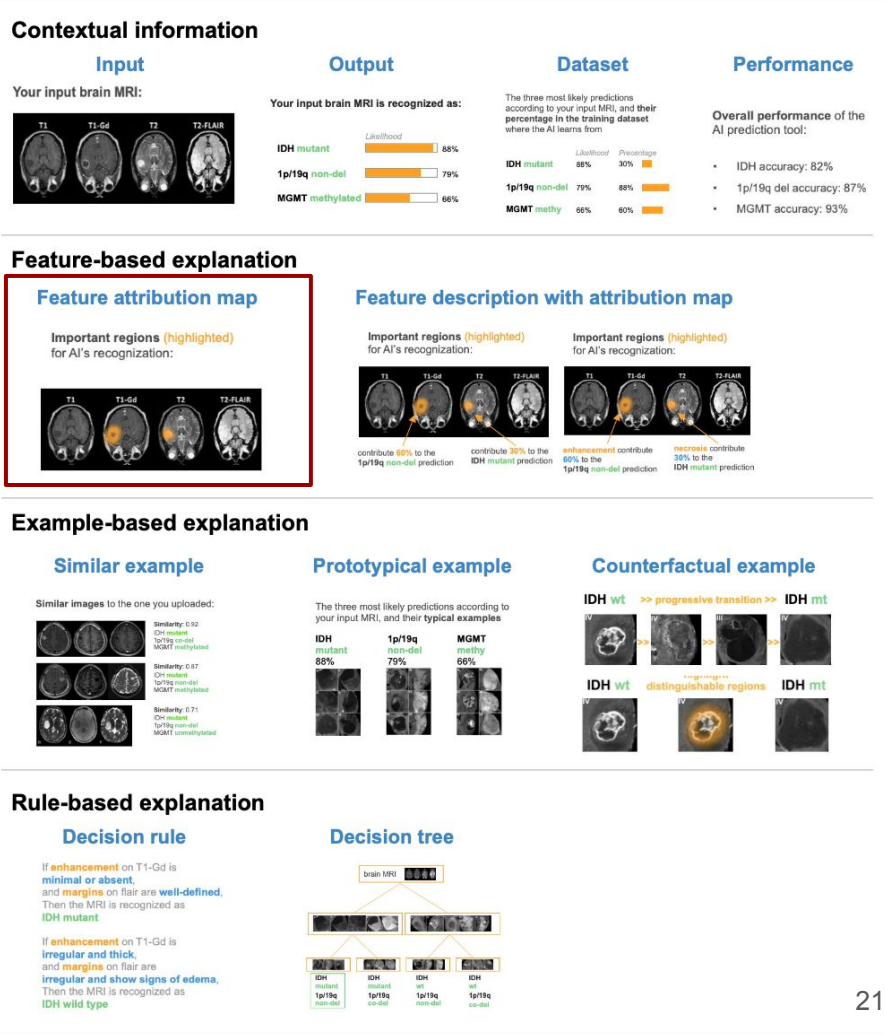




- Card-based low-fidelity prototype result

Co-design XAI with doctors





Core questions

1. What are end user-centered requirements for explainable AI (XAI) algorithms?
2. How to design and evaluate clinical user-centered explainable AI (XAI)?

My user-centered approach

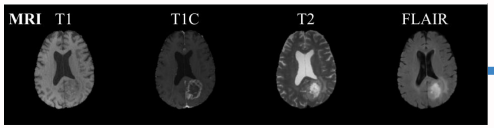
1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors



- A functioning prototype

Clinical decision support using AI model

Clinical task: brain tumor grading task: recognize the brain tumor as *low* or *high* grade



Input MRI

AI prototype

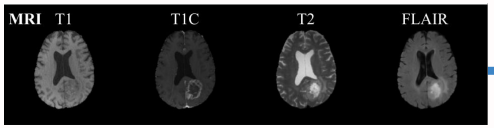
AI's suggestion:
Grade 4 glioblastoma

Accuracy: 88%

- A functioning prototype

Clinical decision support using AI model

Clinical task: brain tumor grading task: recognize the brain tumor as *low* or *high* grade



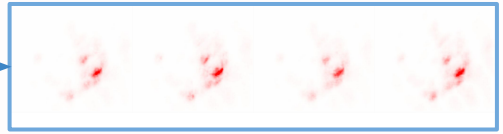
Input MRI

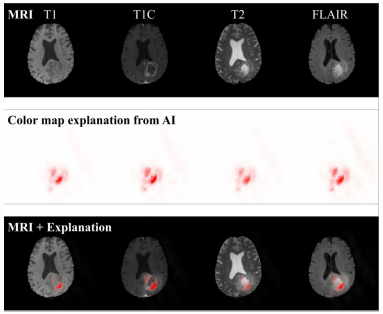
AI prototype

AI's suggestion:
Grade 4 glioblastoma

Accuracy: 88%

AI's explanation:





Feature attribution explanation

Core questions

1. What are end user-centered requirements for explainable AI (XAI) algorithms?
2. How to design and evaluate clinical user-centered explainable AI (XAI)?

My user-centered approach

1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors



Potential clinical utilities of AI explanations to doctors

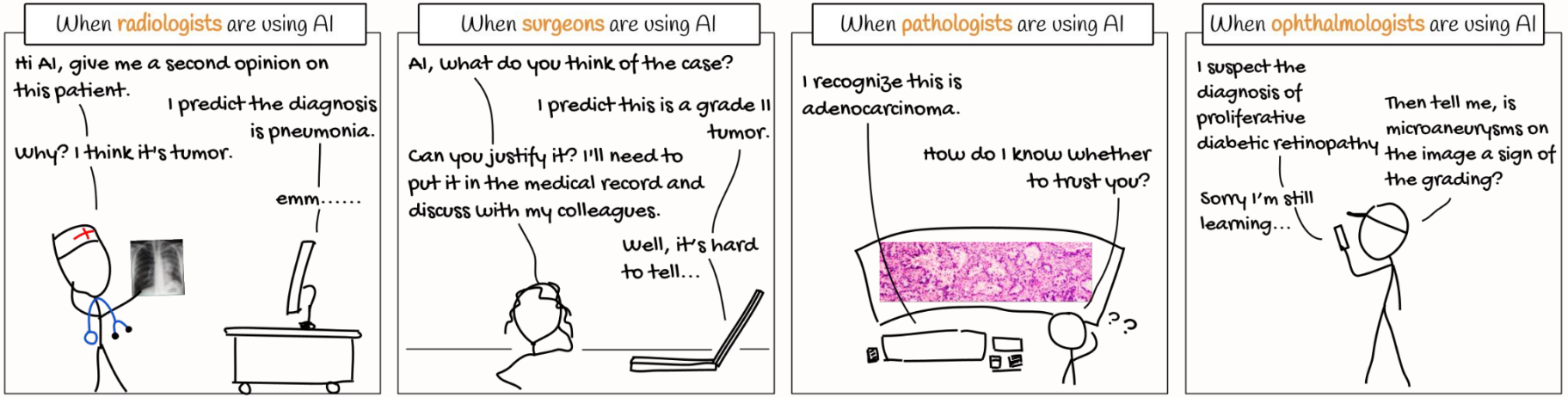
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

Resolve clinical users' disagreement with AI [1]
informed decision

Verify AI's suggestion [1]
evidence-based medicine

Calibrate clinical users' trust [1]

Learning & medical knowledge discovery [1]



[1] EUCA: the End-User-Centered Explainable AI Prototyping Framework. arXiv:2102.02437
[2] Towards A Rigorous Science of Interpretable Machine Learning

- A mixed-method user study with doctors

Potential clinical utilities of AI explanations to doctors

Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

Resolve clinical users' disagreement with AI [1]
informed decision

Verify AI's suggestion [1]
evidence-based medicine

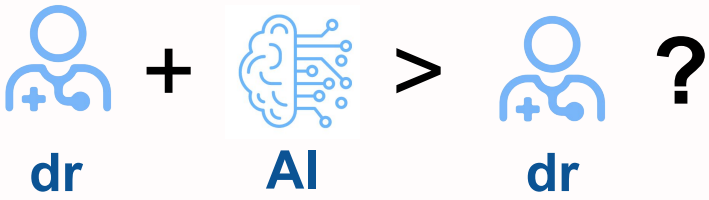
Calibrate clinical users' trust [1]

Learning & medical knowledge discovery [1]

Research question:

Is the AI assistance (prediction + explanation) helpful for clinical users?

Hypothesis:



Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- A mixed-method user study with doctors

Study design

Each doctor underwent three conditions

Measurement: task performance



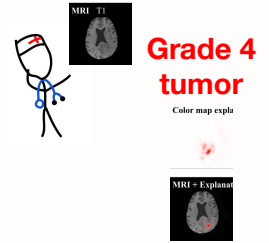
1. **DR**

Doctor only



2. **DR+AI**

Doctor assisted by AI **prediction**



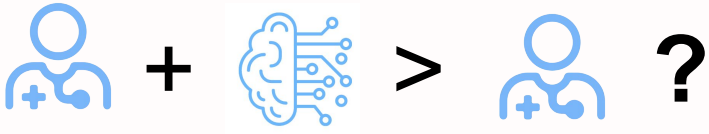
3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**

Research question:

Is the AI assistance (prediction + explanation) helpful for clinical users?

Hypothesis:



prediction, explanation

Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- A mixed-method user study with doctors

This healthcare case is relevant to similar issues when using AI to assist users

Students using AI,

HR using AI,

Lawyers using AI,

Loan officer using AI,

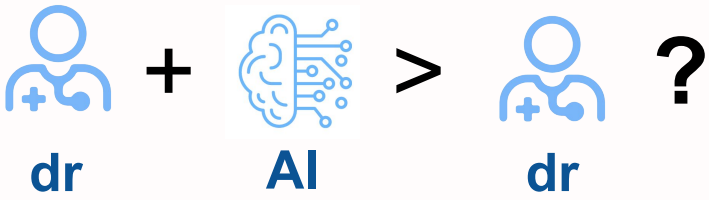
Drivers using AI,

.....

Research question:

Is the AI assistance (prediction + explanation) helpful for clinical users?

Hypothesis:



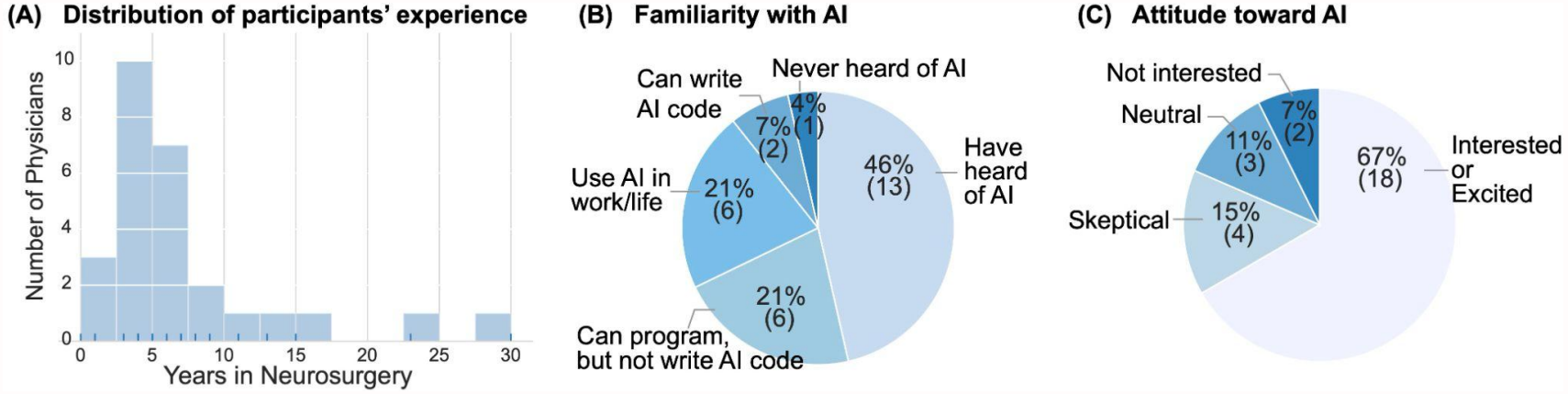
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

● A mixed-method user study with doctors

A user study on the clinical utility of AI and its explanation in a brain tumor grading task

Study Participants:

- 35 neurosurgeons
- 12 attending, 2 fellows, 21 residents
- Years of practicing neurosurgery: 7.1 ± 6.5
- Female: male = 7:19; age: 34.7±8.2



Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task.
W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- **A mixed-method user study with doctors**

Study procedure

35 neurosurgeons, each read 25 brain images

Drs gave judgment on three conditions:



1. **DR**

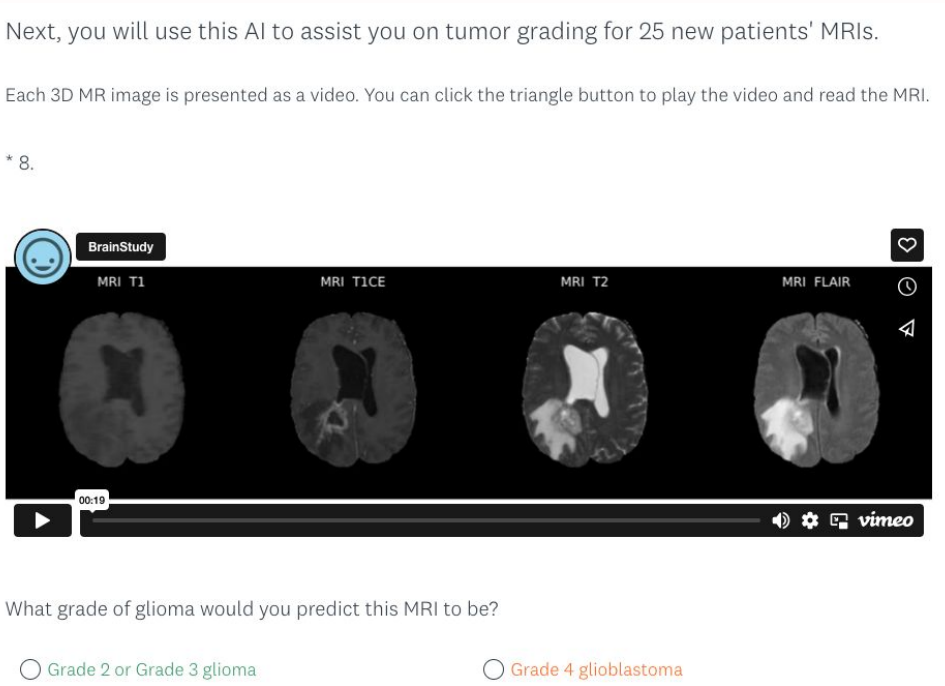
Doctor only

2. **DR+AI**

Doctor assisted by AI **prediction**

3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**



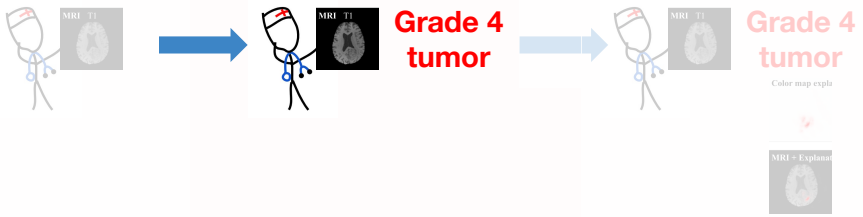
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- A mixed-method user study with doctors

Study procedure

35 neurosurgeons, each read 25 brain images

Drs gave judgment on three conditions:



1. DR

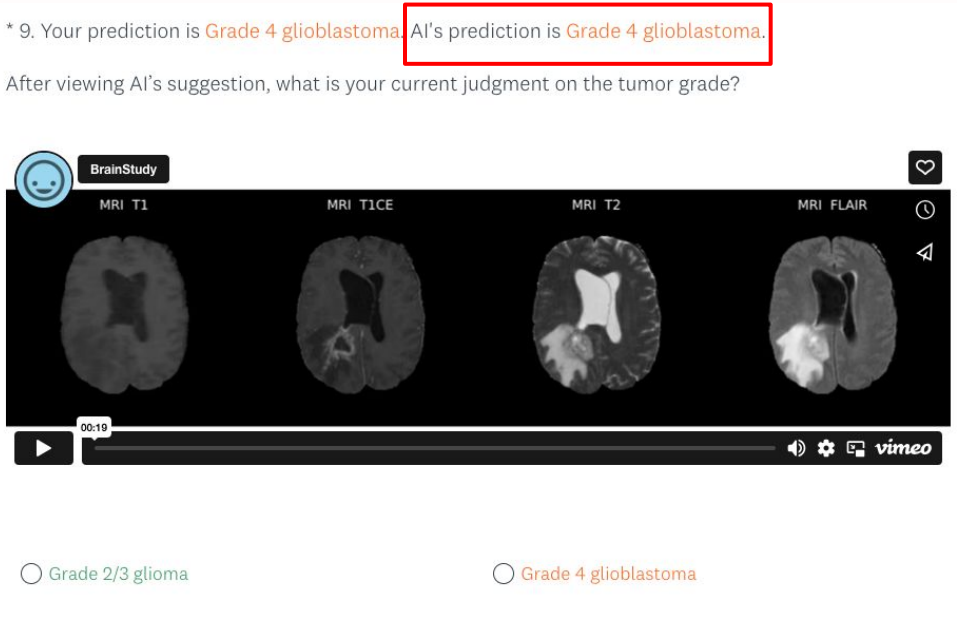
Doctor only

2. **DR+AI**

Doctor assisted by AI **prediction**

3. DR+XAI

Doctor assisted by AI prediction & explanation



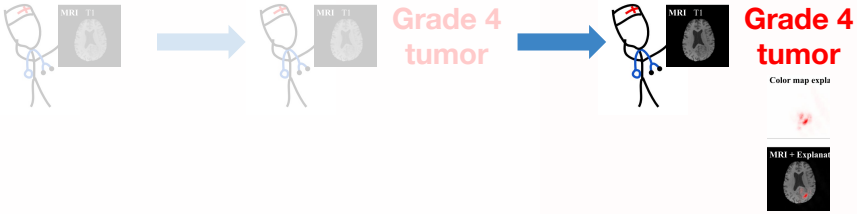
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- **A mixed-method user study with doctors**

Study procedure

35 neurosurgeons, each read 25 brain images

Drs gave judgment on three conditions:



1. DR

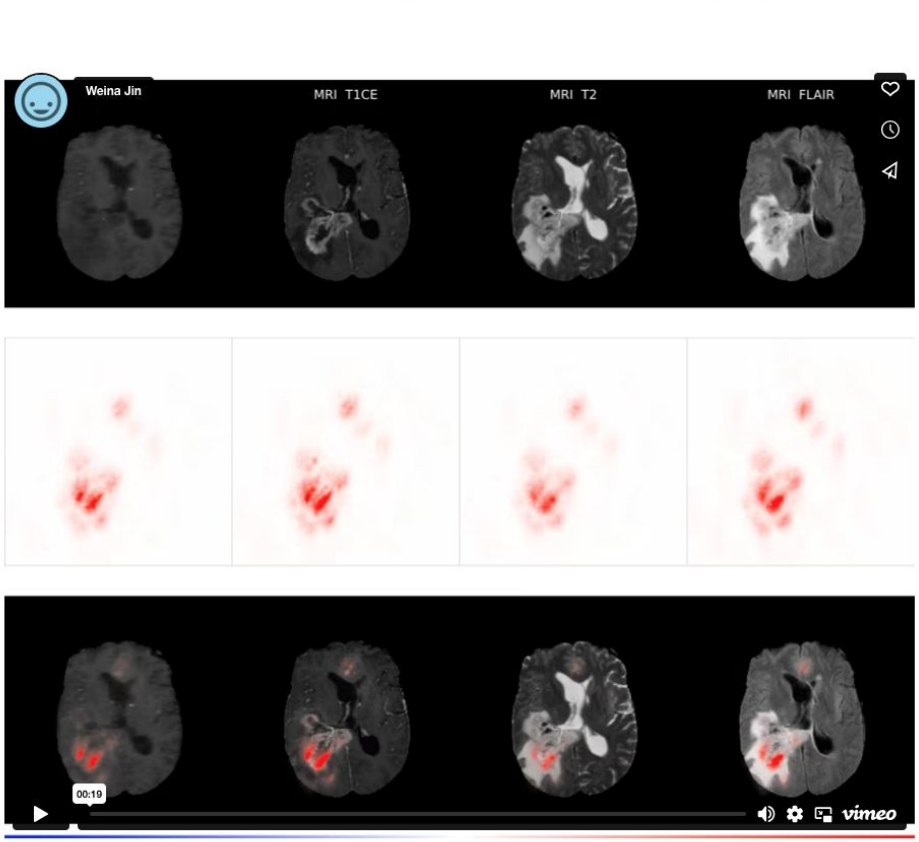
Doctor only

2. DR+AI

Doctor assisted by AI prediction

3. **DR+XAI**

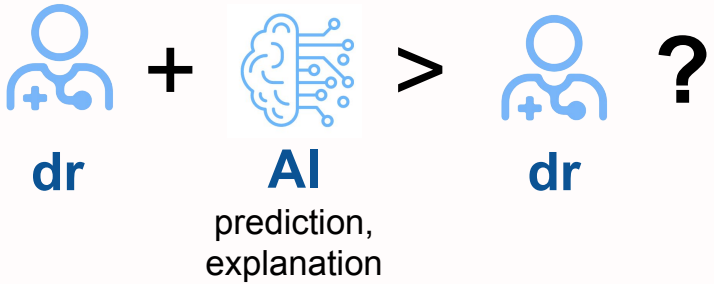
Doctor assisted by AI prediction & **explanation**

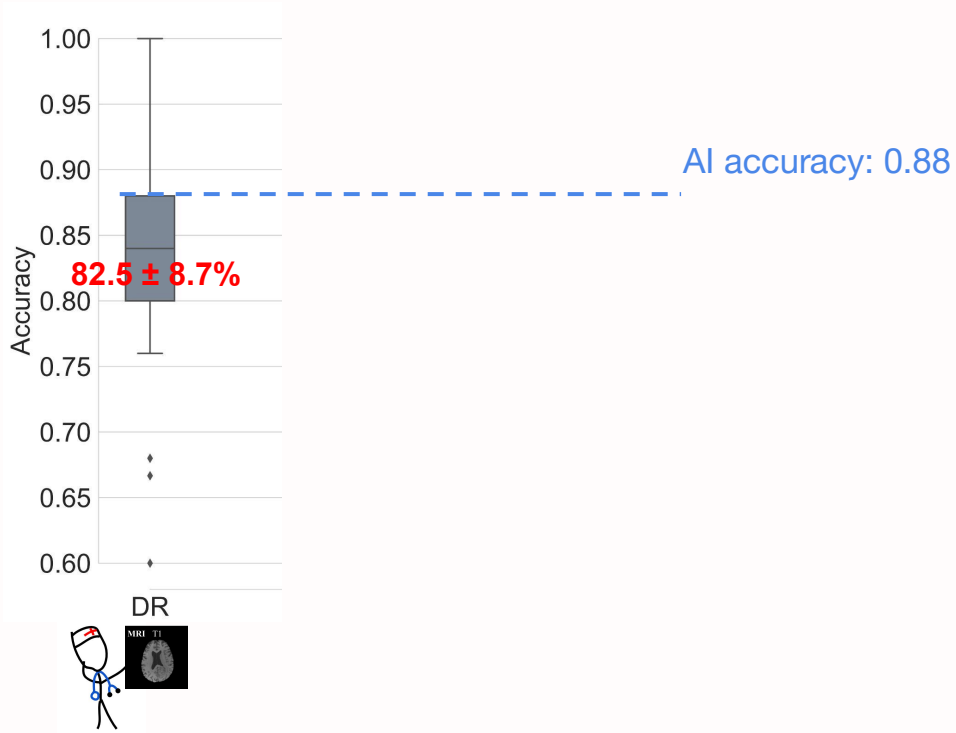


- A mixed-method user study with doctors

Study result

Is doctor+AI better than doctor alone?



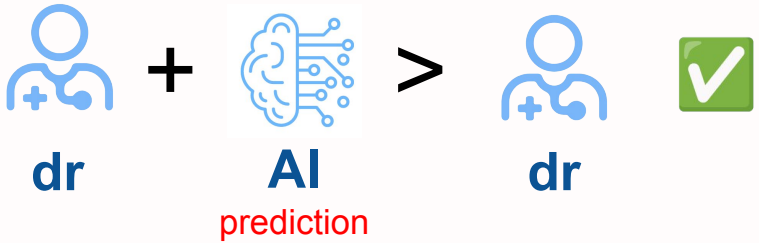


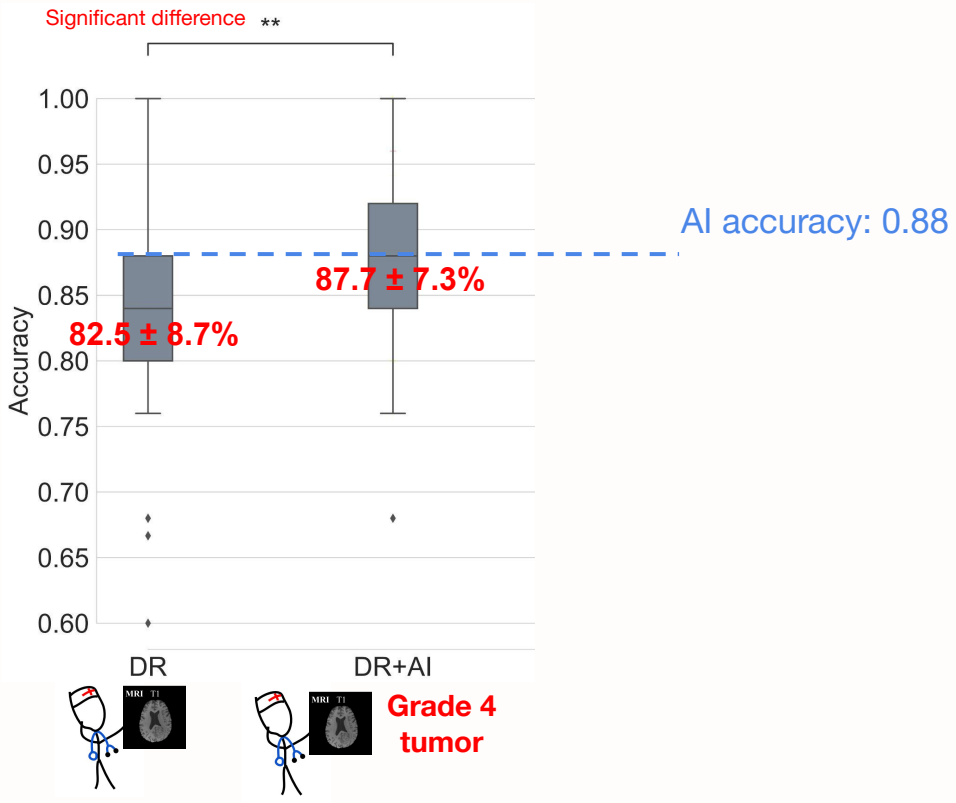
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- **A mixed-method user study with doctors**

Study result

Is doctor+AI better than doctor alone?





In an experimental setting:

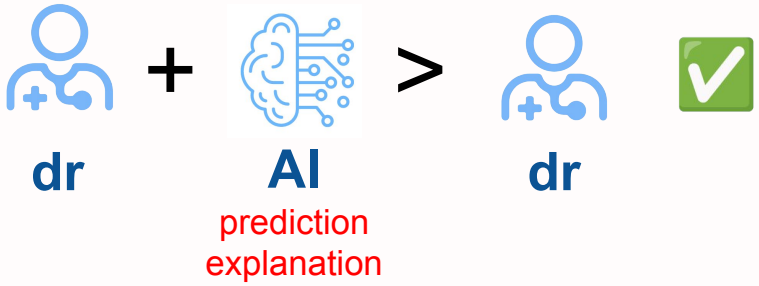
- AI prediction (**DR+AI**) is helpful to improve doctors' task performance

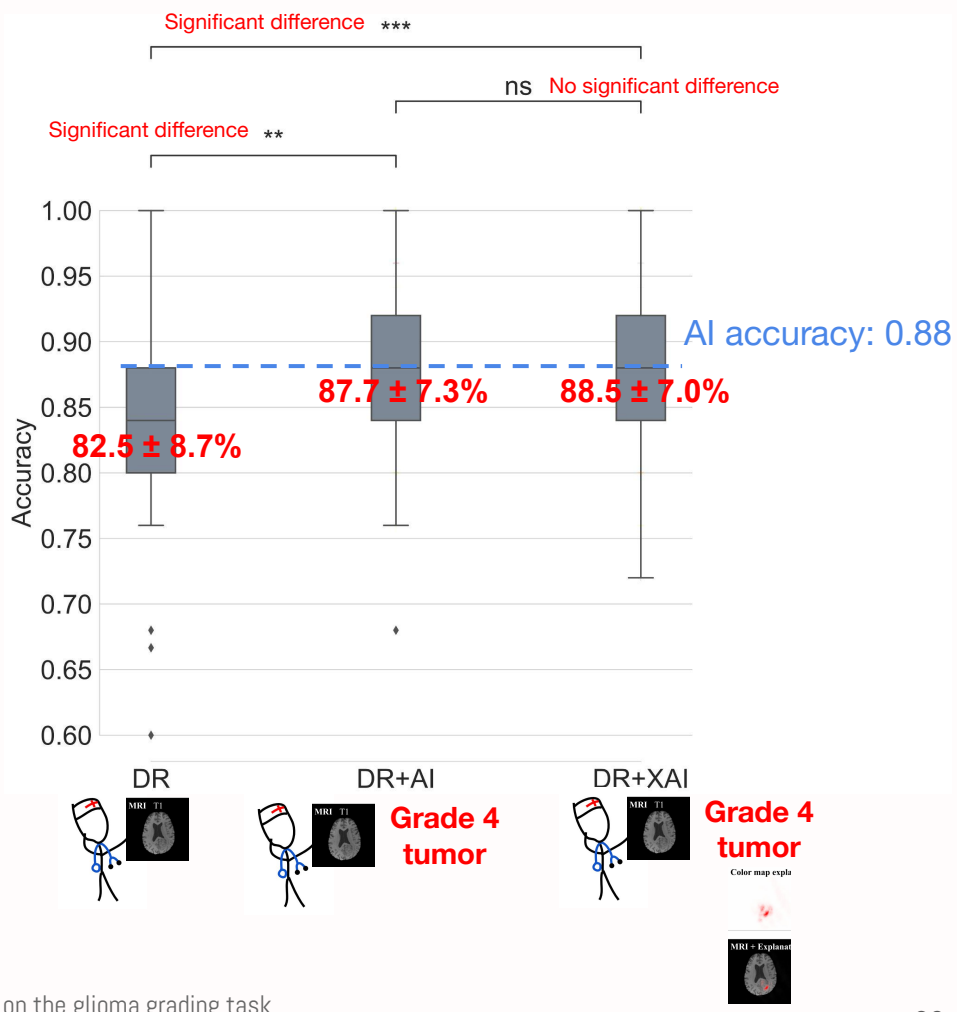
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

- A mixed-method user study with doctors

Study result

Is doctor+AI better than doctor alone?





In an experimental setting:

- AI prediction (**DR+AI**) is helpful to improve doctors' task performance
- AI explanation (**DR+XAI**) does not show additional help in improving doctors' performance

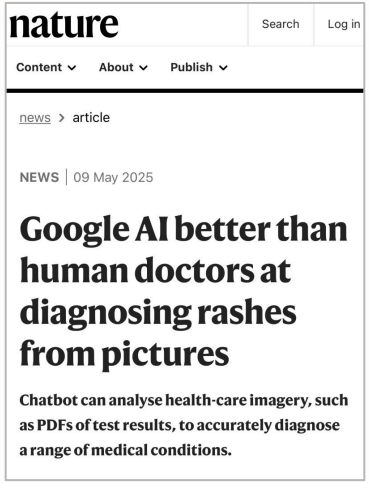
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task.

This exercise is relevant to similar issues when using AI to assist users

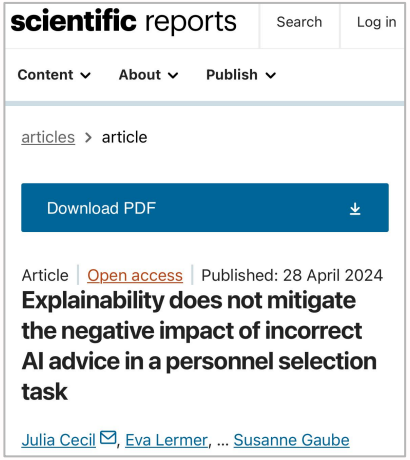
Similar findings and claims on

AI outperforms humans in a, b, c, … tasks

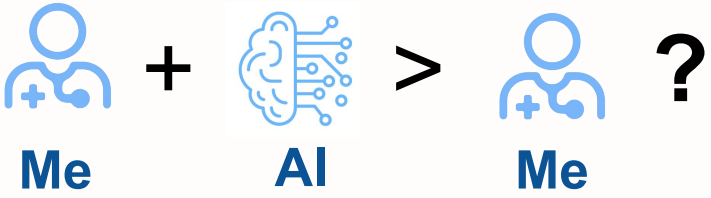
AI explanation fails in x, y, z, … tasks











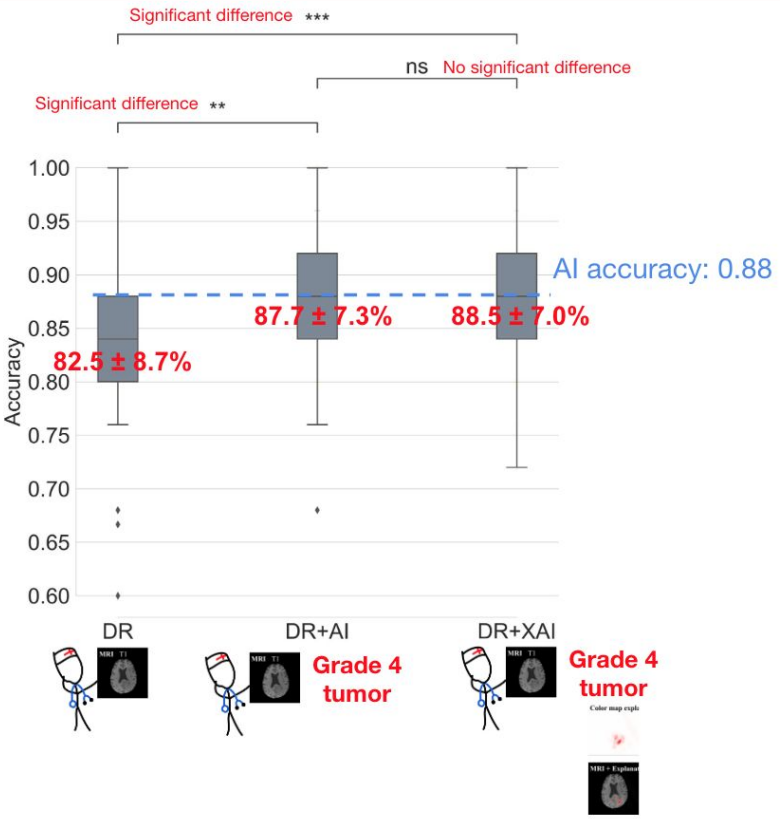
prediction,explanation

- A mixed-method user study with doctors

We found the same evidence can be interpreted differently to support opposite recommendations

AI optimism

1. Black-box AI is **superior** to doctors;
2. Explainable AI is **useless**.
3. So we can **replace doctors with black-box AI** to improve medical decisions.



Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

- A mixed-method user study with doctors

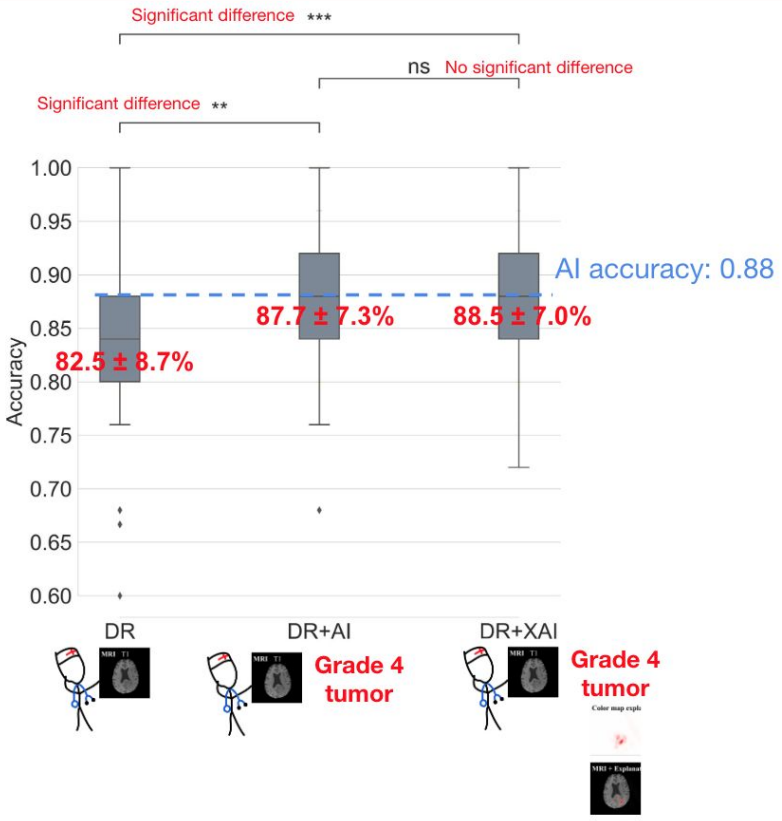
We found the same evidence can be interpreted differently to support opposite recommendations

AI optimism

1. Black-box AI is **superior** to doctors;
2. Explainable AI is **useless**.
3. So we can **replace doctors with black-box AI** to improve medical decisions.

AI skepticism

1. This is a **small sample** study in an experimental setting. The result needs to be **further verified** in large-scale, real-world tasks.
2. Other factors can influence/explain away the results, such as **overreliance** on AI, and **weakness in XAI**.



Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

- A mixed-method user study with doctors

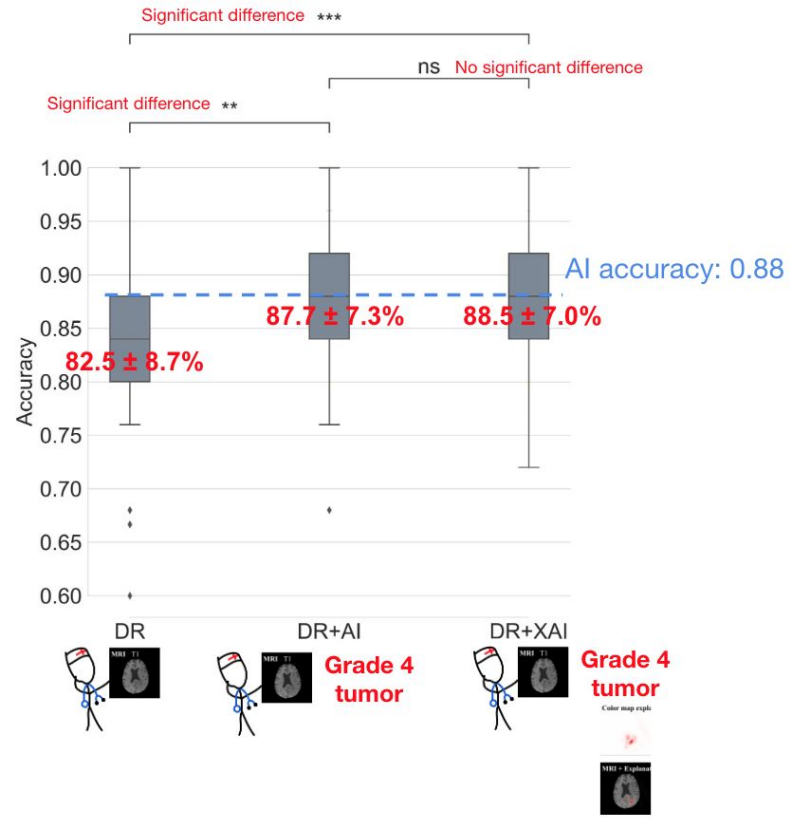
We found the same evidence can be interpreted differently to support opposite recommendations

AI optimism → Develop AI to **compete** with humans

1. Black-box AI is **superior** to doctors;
2. Explainable AI is **useless**.
3. So we can **replace doctors with black-box AI** to improve medical decisions.

AI skepticism → Develop AI to **collaborate** with humans

1. This is a **small sample** study in an experimental setting. The result needs to be **further verified** in large-scale, real-world tasks.
2. Other factors can influence/explain away the results, such as **overreliance** on AI, and **weakness in XAI**.



Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

- A mixed-method user study with doctors

We found the same evidence can be interpreted differently to support opposite recommendations

AI optimism → Develop AI to **compete** with humans

1. Black-box AI is **superior** to doctors;
2. Explainable AI is **useless**.
3. So we can **replace doctors with black-box AI** to improve medical decisions.

AI skepticism → Develop AI to **collaborate** with humans

1. This is a **small sample** study in an experimental setting. The result needs to be **further verified** in large-scale, real-world tasks.
2. Other factors can influence/explain away the results, such as **overreliance** on AI, and **weakness in XAI**.

Critical thinking requires us to have more prudence and holdback from claims like this one that may create significant impact.

Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

- A mixed-method user study with doctors

We found the same evidence can be interpreted differently to support opposite recommendations

AI optimism → Develop AI to **compete** with humans

1. Black-box AI is **superior** to doctors;
2. Explainable AI is **useless**.
3. So we can **replace doctors with black-box AI** to improve medical decisions.



Critical thinking requires us to have more prudence and holdback from claims like this one that may create significant impact.

AI skepticism → Develop AI to **collaborate** with humans

1. This is a **small sample** study in an experimental setting. The result needs to be **further verified** in large-scale, real-world tasks.
2. Other factors can influence/explain away the results, such as **overreliance** on AI, and **weakness in XAI**.



That means we need to carefully consider the underlying assumptions and any possible counter arguments and counter evidence to them, especially from people with diverse perspectives and interests.

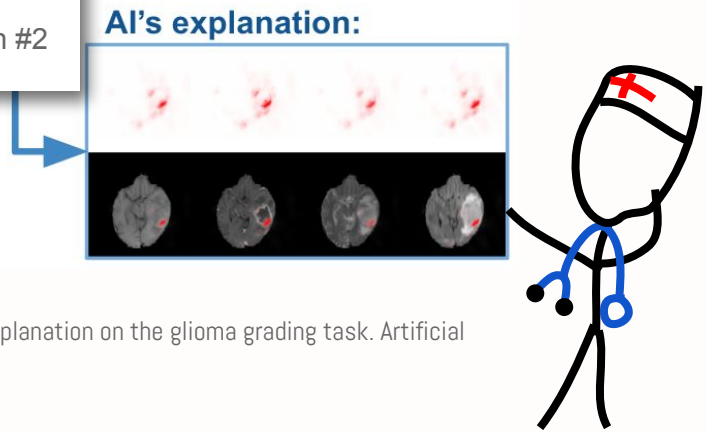
Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

● Qualitative study result

Feature attribution explanation, the most commonly used explanation form, is not useful for doctors

> Though the color map is drawing your eyes to many different spots, but I feel like I didn't understand why my eyes were being driven to those spots, like why were these very specific components important? And I think that's where all **my confusion** was.
>
> – Neurosurgeon #2

> **What does that (color map region) mean?** Like hey, which part of my car gets my car moving? It should say press the accelerator. But yours would just show a dashboard of the car, and show that this button had some red, that button had some red, but it's not an explanation. Let's go to an example, and you'll see, what about the red areas under MRI T1CE (modality)? **Was it central necrosis?** But it couldn't be the central necrosis, because there's more central necrosis in the temporal lobe, and that area didn't get highlighted. So anyway, I don't know, it's just **confusing**.
>
> …These color maps were **totally useless** **without text, without any context or explanation**, like those details. The color maps were just pretty, but they didn't explain anything.
>
> – Neurosurgeon #3



Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. Artificial Intelligence in Medicine. 2024

XAI designed for technical users' toy tasks



Natural image recognition

- Trivial and low-stakes usage scenarios
- Do not require explanation

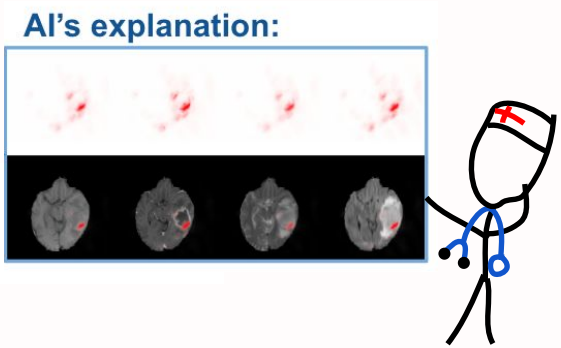
Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

XAI designed for technical users' toy tasks is not useful for end-users' real-world tasks



Natural image recognition

- Trivial and low-stakes usage scenarios
- Do not require explanation



Medical image recognition

- High-stakes, complex usage scenarios
- Explainability is originally motivated

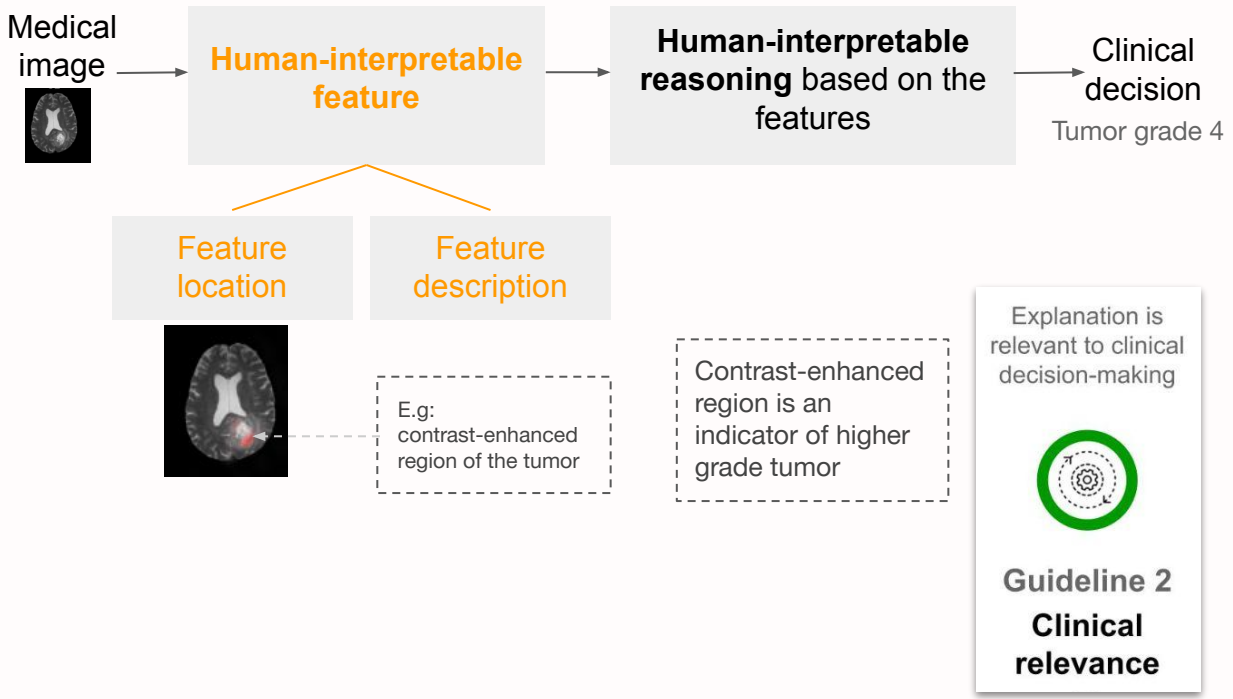
Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

● Qualitative study result

The explanation form should align with clinical reasoning and explanatory process

> 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 explanation process:



Core questions

1. **What are end user-centered requirements for explainable AI (XAI) algorithms?**
2. How to design and evaluate clinical user-centered explainable AI (XAI)?



My user-centered approach

1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors



W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023

Core questions

1. **What are end user-centered requirements for explainable AI (XAI) algorithms?**
2. How to design and evaluate clinical user-centered explainable AI (XAI)?

My user-centered approach

1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors

1. The user study that is meant to use AI to assist users can be interpreted as evidence to replace users.
2. Feature attribution explanation, the most commonly used explanation form, is not useful for doctors.



Core questions

1. What are end user-centered requirements for explainable AI (XAI) algorithms?
2. How to design and evaluate clinical user-centered explainable AI (XAI)?

My user-centered approach

1. Usability evaluation
2. Low-fidelity prototype
3. A pilot user study with doctors
4. Functioning prototype
5. A mixed-method user study with doctors

1. The user study that is meant to use AI to assist users can be interpreted as evidence to replace users.
2. Feature attribution explanation, the most commonly used explanation form, is not useful for doctors.

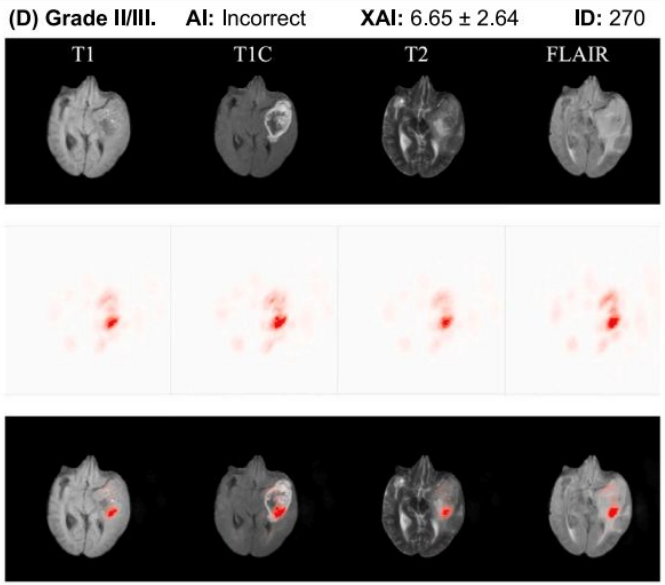


● An observation from the qualitative user study

Talking with doctors reveals a problem in the XAI evaluation paradigm

> If a system made its prediction based upon these areas (outside the tumor), I would definitely not trust that system, but I would be very reassured that the system is telling me that.
> ...You show me this (color map), and I see that it's focusing a lot of its decision-making on areas that have nothing to do with the tumor. That decreases the trust in that specific model (AI decision on this specific case), but it increases the overall trust in AI. So I'm less likely to use this model, but I'm more likely to use a model that does a better job than this, because I am reassured that when I see that better model, that I will be able to have access to that back-end explanation. So you're not asking me to believe in the system without evidence.
>
> – Neurosurgeon #1

This "better", more reasonable explanation explains a **wrong** prediction from AI.



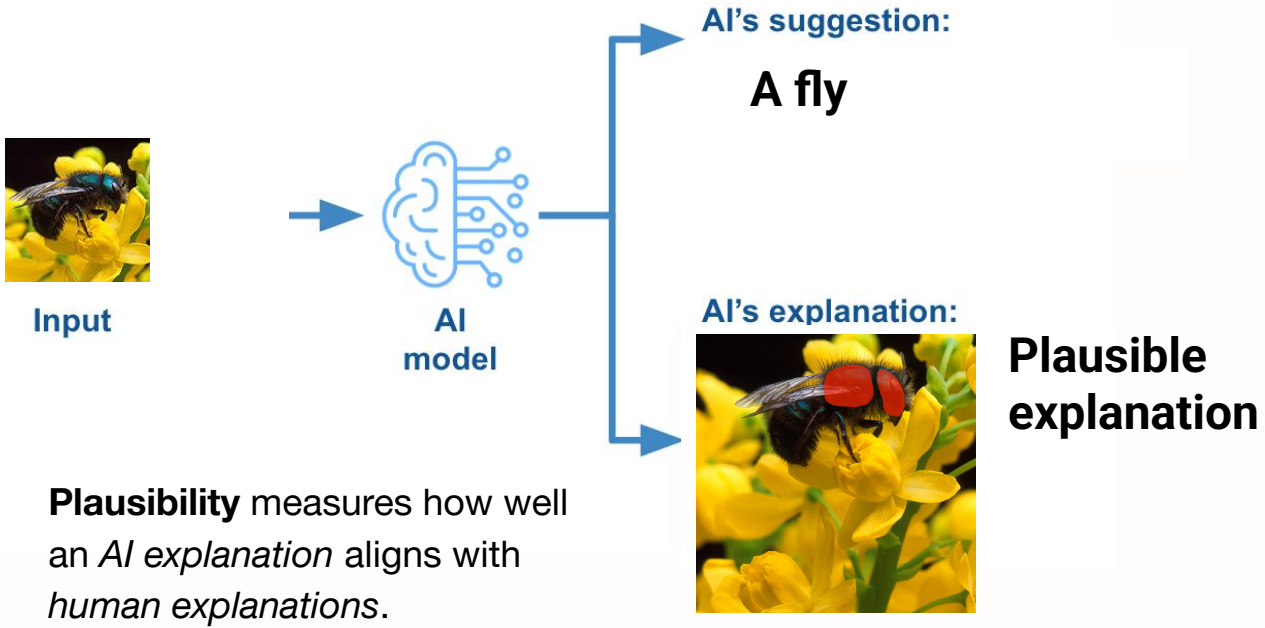
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

Plausibility, the most commonly used XAI evaluation metric [1]

Plausibility measures how well an *AI explanation* aligns with *human explanations*.

[1] From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys. 2023
[2] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Plausibility, the most commonly used XAI evaluation metric [1]



Plausibility measures how well an *AI explanation* aligns with *human explanations*.

[1] From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys. 2023
[2] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Plausibility, the most commonly used XAI evaluation metric [1], can mislead and manipulate end-users [2]



Input



AI model

AI's suggestion:

~~A fly~~

A bee

Wrong prediction

AI's explanation:



Plausible explanation

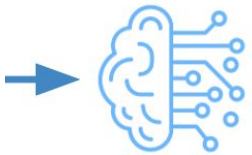
Plausibility measures how well an *AI explanation* aligns with *human explanations*.

[1] From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys. 2023
[2] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Plausibility, the most commonly used XAI evaluation metric [1], can mislead and manipulate end-users [2]



Input



AI model

AI's suggestion:

~~A fly~~
A bee

Wrong prediction

+

AI's explanation:



Plausible explanation

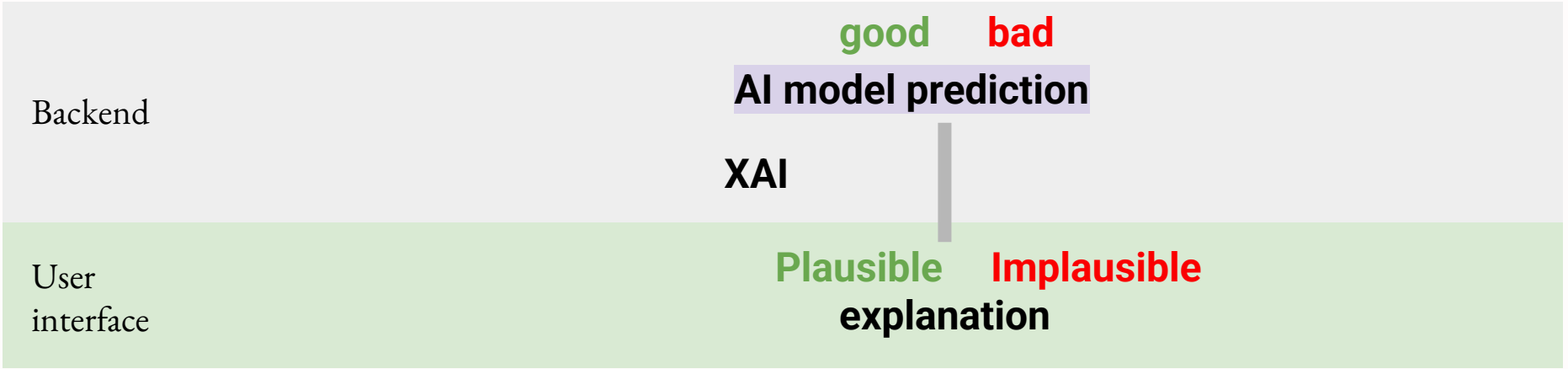
=

Misleading explanation

Plausibility measures how well an *AI explanation* aligns with *human explanations*.

[1] From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Computing Surveys. 2023
[2] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Sometimes, pursuing user preference/satisfaction may go against users' own interests and benefits

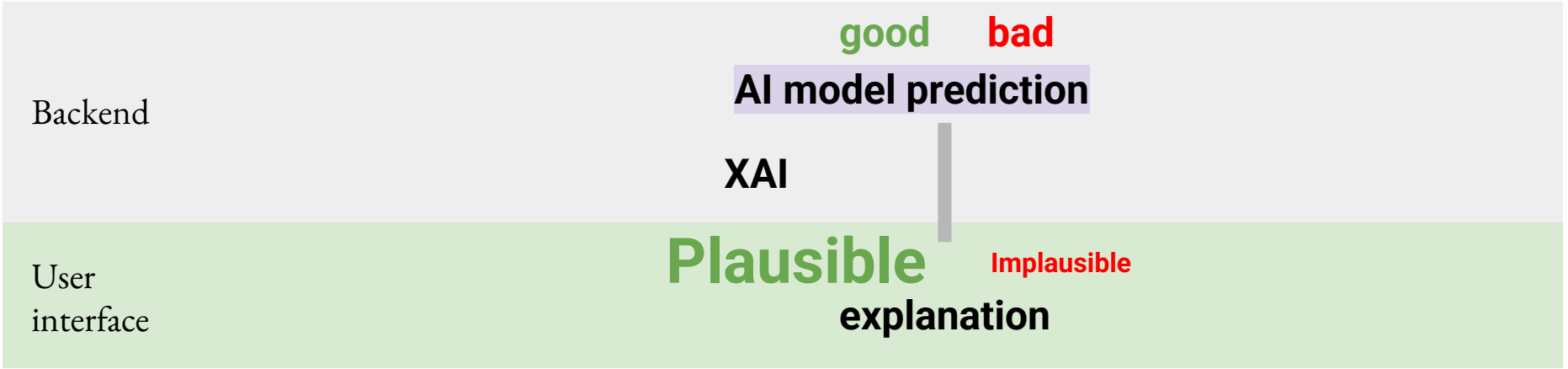


Plausibility measures how well an *AI explanation* aligns with *human explanations*.

⇨ Optimize/evaluate XAI using plausibility ⇨ We want more plausible explanations

[1] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Sometimes, pursuing user preference/satisfaction may go against users' own interests and benefits

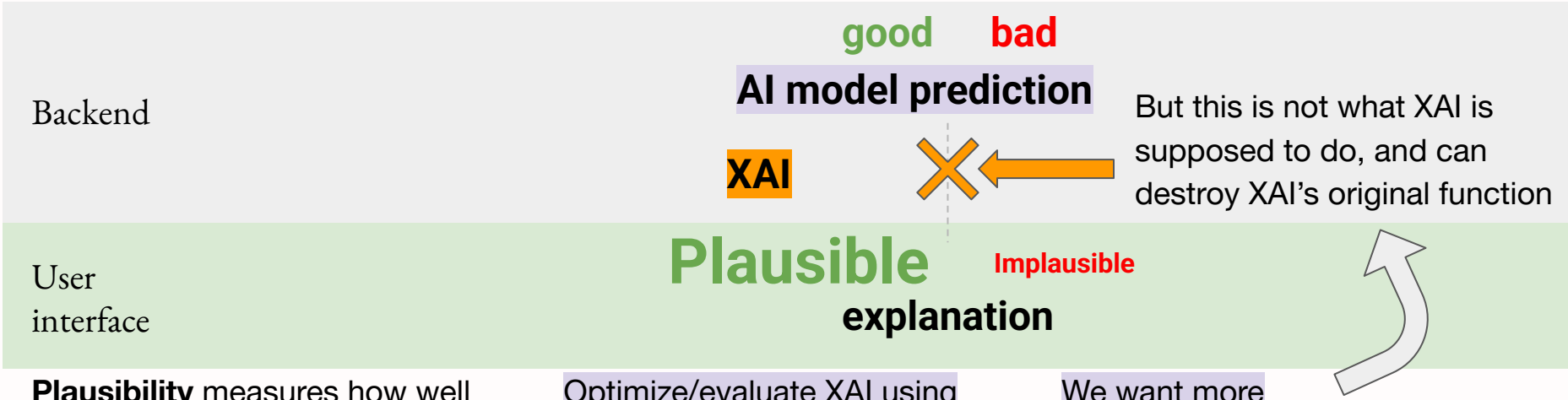


Plausibility measures how well an *AI explanation* aligns with *human explanations*.

⇨ Optimize/evaluate XAI using plausibility ⇨ We want more plausible explanations

[1] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Sometimes, pursuing user preference/satisfaction may go against users' own interests and benefits



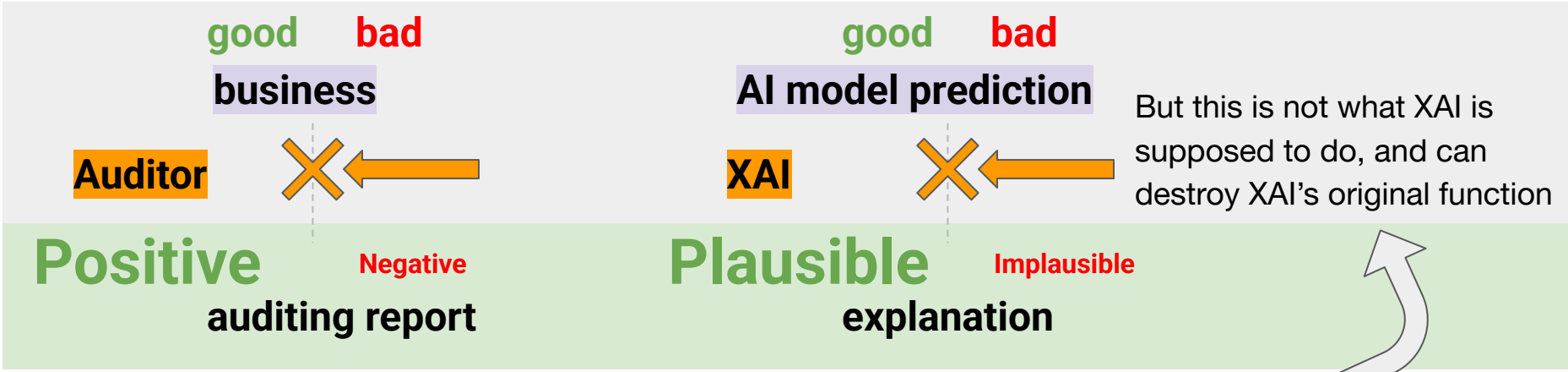
Plausibility measures how well an *AI explanation* aligns with *human explanations*.

Optimize/evaluate XAI using plausibility

We want more plausible explanations

[1] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Sometimes, pursuing user preference/satisfaction may go against users' own interests and benefits



Plausibility measures how well an *AI explanation* aligns with *human explanations*.

Optimize/evaluate XAI using plausibility

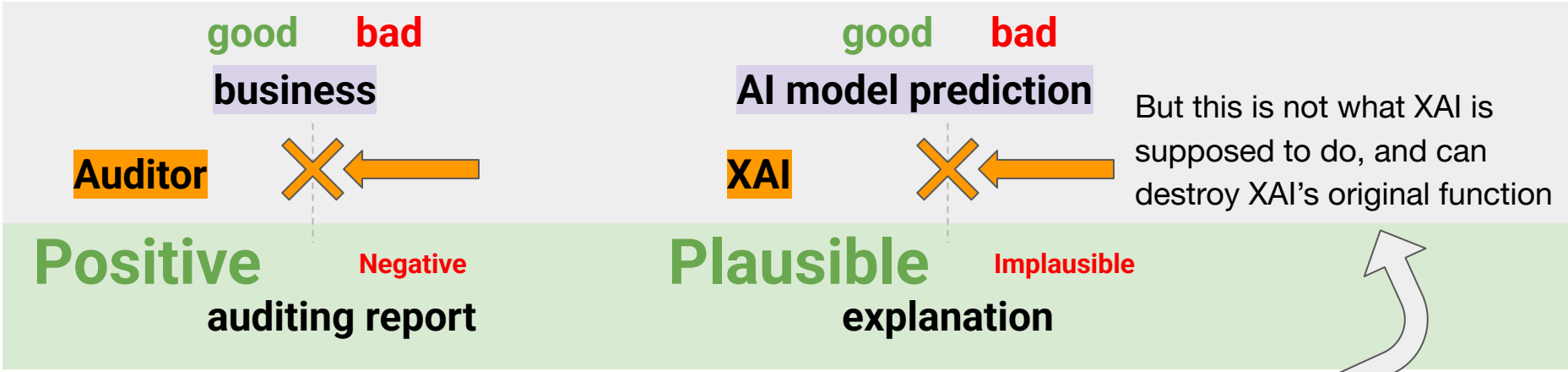
We want more plausible explanations

[1] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Sometimes, pursuing user preference/satisfaction may go against users' own interests and benefits

Incentivize (bribe) auditors to generate more positive reports

Incentivize (bribe) XAI algorithms to generate more plausible explanations



Plausibility measures how well an *AI explanation* aligns with *human explanations*.

⇨ Optimize/evaluate XAI using plausibility ⇨ We want more plausible explanations

[1] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

- Critically examine the human-centered concept

What does “user-centered” really mean?

1. The user study that is meant to use AI to assist users can be interpreted as evidence to replace users.
2. The most popular XAI algorithm is not useful for end users.
3. The most popular evaluation metric used in XAI can mislead and manipulate users.

Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

- Critically examine the human-centered concept

What does “user-centered” really mean?

1. The user study that is meant to use AI to assist users can be interpreted as evidence to replace users.
2. The most popular XAI algorithm is not useful for end users.
3. The most popular evaluation metric used in XAI can mislead and manipulate users.

While these findings show problems in user-centered XAI,

these problems are identified through user studies.

Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

- Critically examine the human-centered concept

What does “user-centered” really mean?



1. The user study that is meant to use AI to assist users can be interpreted as evidence to replace users.
2. The most popular XAI algorithm is not useful for end users.
3. The most popular evaluation metric used in XAI can mislead and manipulate users.

User-centered method is a tool:

- Can be captured by big techs for branding, ethics washing, user manipulation and persuasion, and shift responsibility from technical sector to users.
- Can also be used to genuinely improve users’ interests and benefits.

We have the responsibility to critically reflect the purposes of user-centered design:

- **by whom?** (community/user lead design, or by corporation/for-profit interests)
- **for what purpose?** (be aware of the implicit user manipulation and persuasion)

Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588

● Critically examine the human-centered concept

What does “user-centered” really mean?



Related works on critical HCI

- Design Justice: Community-Led Practices to Build the Worlds We Need
- The Dark (Patterns) Side of UX Design
- What is "critical" about critical design?
- Towards a feminist HCI methodology: social science, feminism, and HCI

User-centered method is a tool:

- Can be captured by big techs for branding, ethics washing, user manipulation and persuasion, and shift responsibility from technical sector to users.
- Can also be used to genuinely improve users’ interests and benefits.

We have the responsibility to critically reflect the purposes of user-centered design:

- **by whom?** (community/user lead design, or by corporation/for-profit interests)
- **for what purpose?** (be aware of the implicit user manipulation and persuasion)

Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588