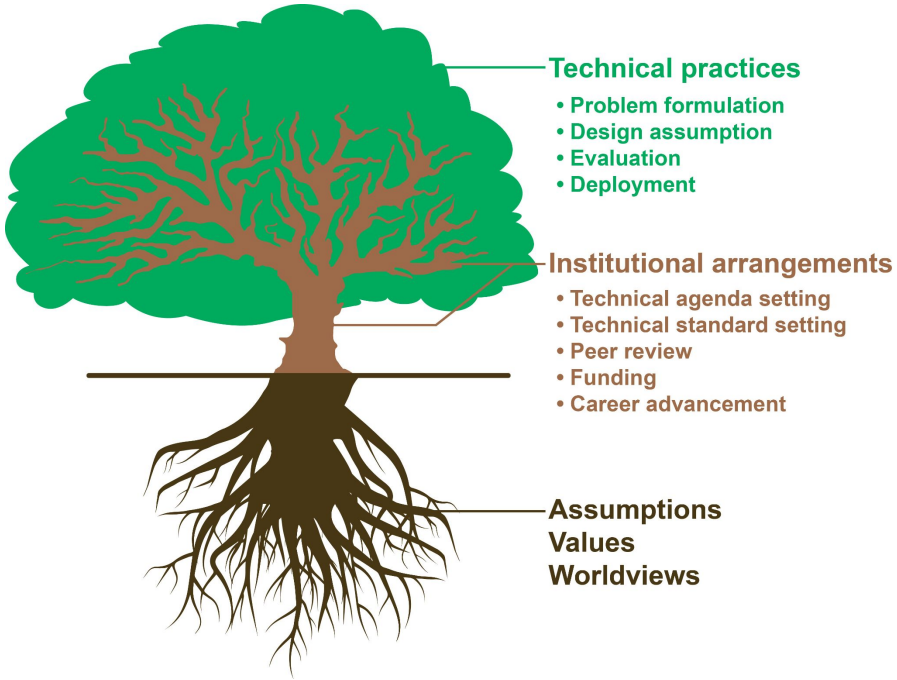
Critical Technical Practice

Motivation and Methods

critical-technical-practice.github.io



Weina Jin, MD

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

weina.me
weinaj@sfu.ca

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
    a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
    a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Critical technical practice

Assumptions

critically inspect assumptions and justifications in common technical practices

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
    a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Critical technical practice

Assumptions critically inspect assumptions and justifications in common technical practices

We cannot make such a certain claim without critical reflections and examinations.

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
    a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Critical technical practice

Assumptions critically inspect assumptions and justifications in common technical practices

We cannot make such a certain claim without critical reflections and examinations.

What if the track/root is wrong? Any post-hoc fix won't help.

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
    a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Critical technical practice

Assumptions critically inspect assumptions and justifications in common technical practices

We cannot make such a certain claim without critical reflections and examinations.

What if the track/root is wrong? Any post-hoc fix won't help.

Is it really natural and inevitable? Are there any alternatives?

My story

5th-year PhD candidate

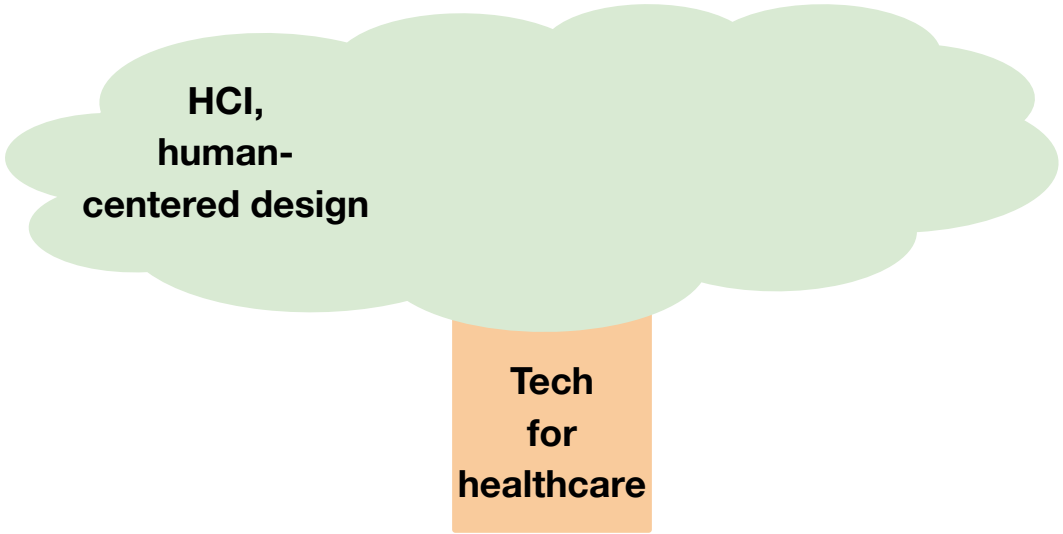
Research interest: how to make AI explainable for clinical users on medical image tasks

Tech for healthcare

My story

5th-year PhD candidate

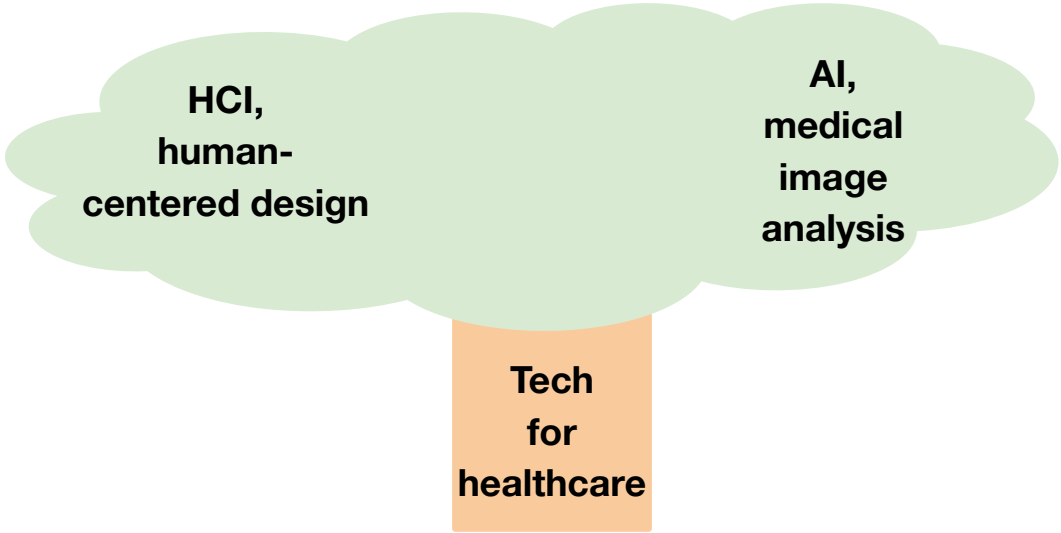
Research interest: how to make AI explainable for clinical users on medical image tasks



My story

5th-year PhD candidate

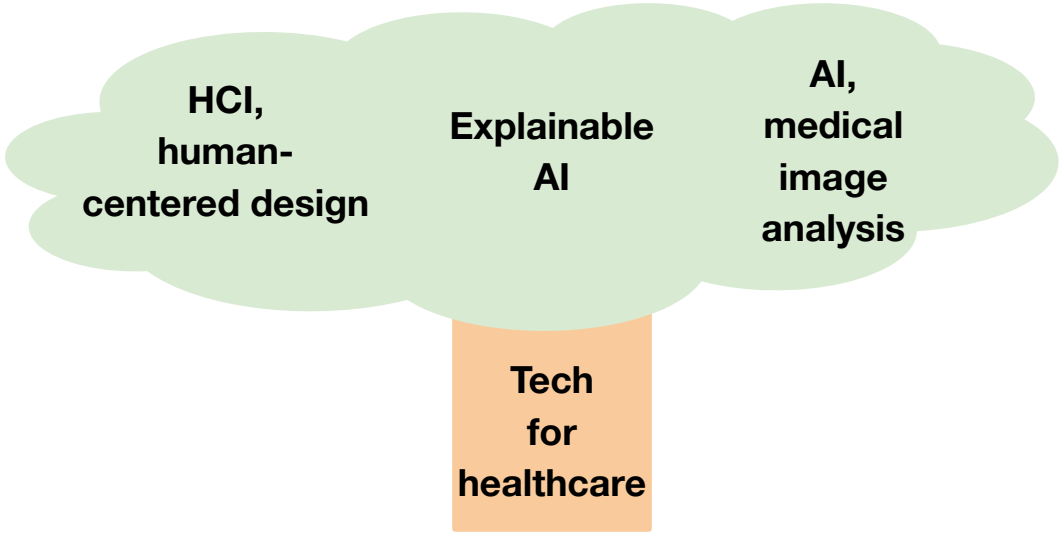
Research interest: how to make AI explainable for clinical users on medical image tasks



My story

5th-year PhD candidate

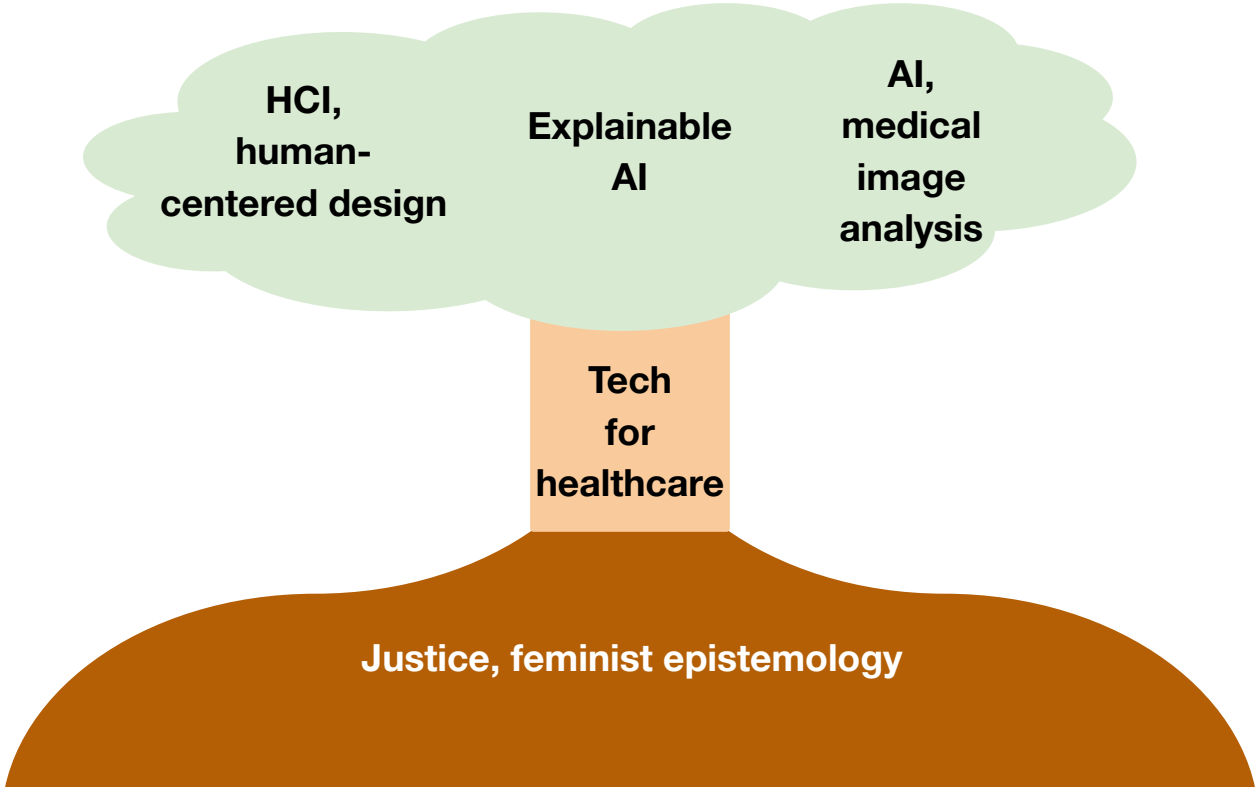
Research interest: how to make AI explainable for clinical users on medical image tasks



My story

5th-year PhD candidate

Research interest: how to make AI explainable for clinical users on medical image tasks



Algorithmic Accountability

Ethical AI

Technology for Good

Diversity

Equity

Fairness

Human-centered technology

AI for Social Good

AI for science

AI Justice

Explainability

AI for Healthcare

Responsible Conduct

Algorithmic Accountability

Ethical AI

Technology for Good

Diversity

Equity

Fairness

Do good intentions in technical development *guarantee* good outcomes?

Human-centered technology

AI for Social Good

AI for science

AI Justice

Explainability

AI for Healthcare

Responsible Conduct

Do good intentions in technical development *guarantee* good outcomes?

Human-centered technology

Explainability AI for Healthcare

Do good intentions in technical development
guarantee
good outcomes?

Human-centered technology

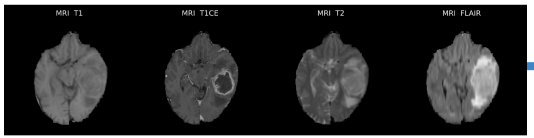
Clinical User-Centred Explainable AI for Medical Image Analysis

Explainability AI for Healthcare

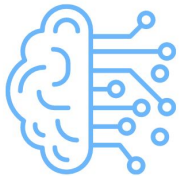
Medical image analysis: the modeling aspect

Use computational methods to interpret and analyze information from medical images

For example, in clinical decision support:



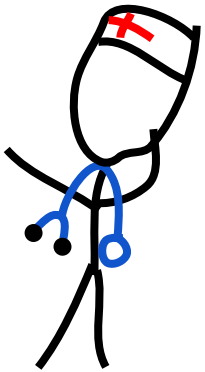
Input brain image



Model

Suggestion from the model:
High-grade brain tumor

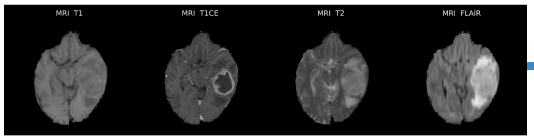
Output



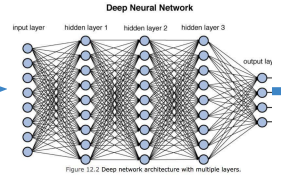
Medical image analysis: the modeling aspect

Use computational methods to interpret and analyze information from medical images

For example, in clinical decision support:



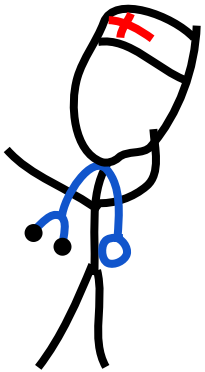
Input brain image



AI model

Suggestion from the model:
High-grade brain tumor

Output



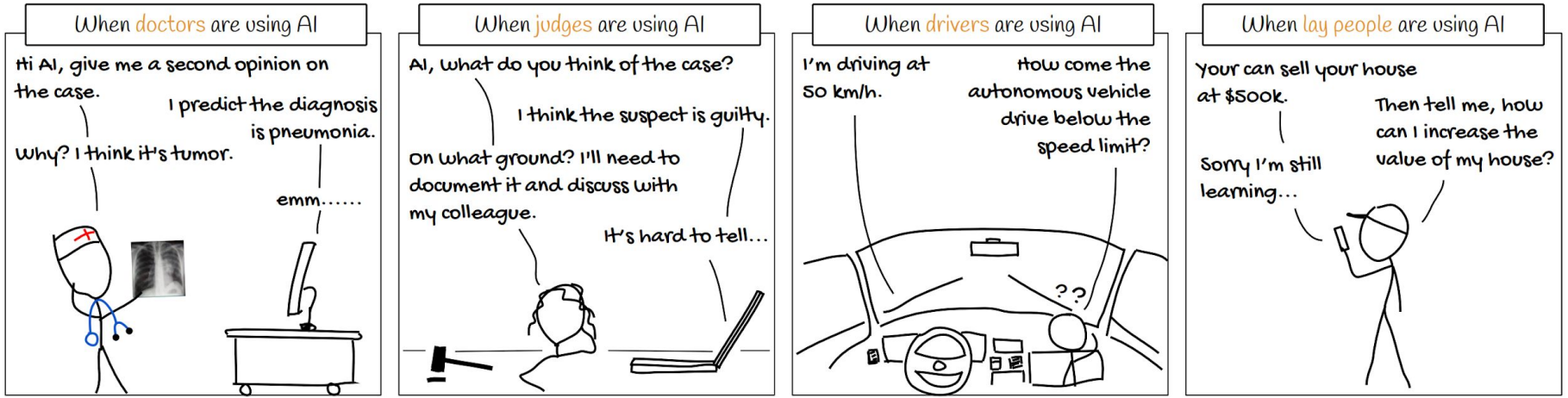
AI models are largely black-box, human-uninterpretable models because of

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

Explainable AI (XAI)

In **critical tasks** that related to humans' life, money, career, etc., for humans to make informed decision assisted by a model,
the model needs to **explain its decisions in human-understandable ways** [1].

Explainability is one of the **AI ethics principles** that enables the other principles [2].



1. Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017
2. Luciano Floridi, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. 2018

Potential clinical utilities of AI explanations: the clinical users' aspect

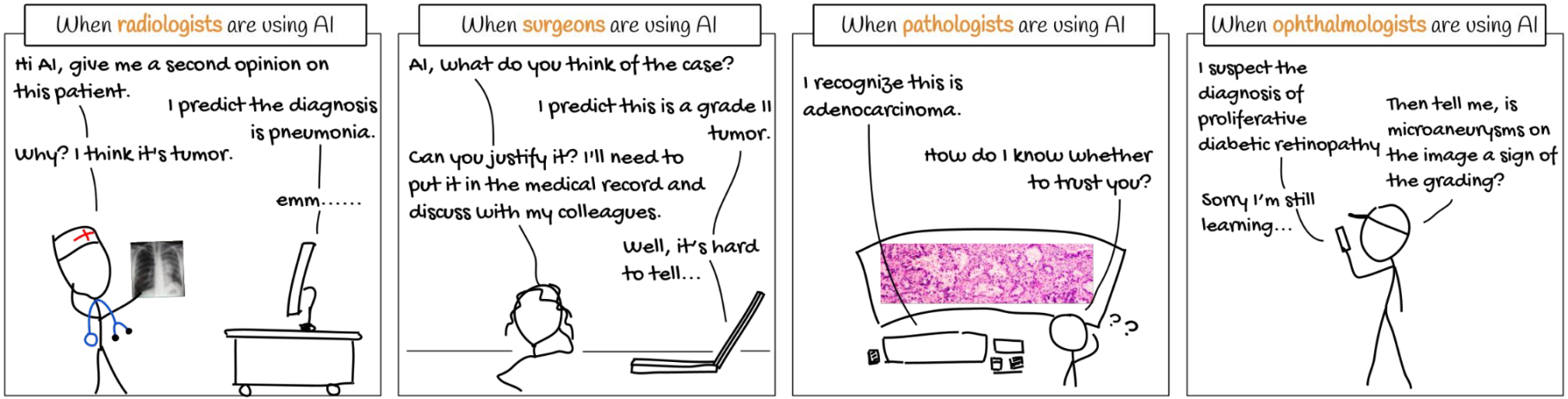
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

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?

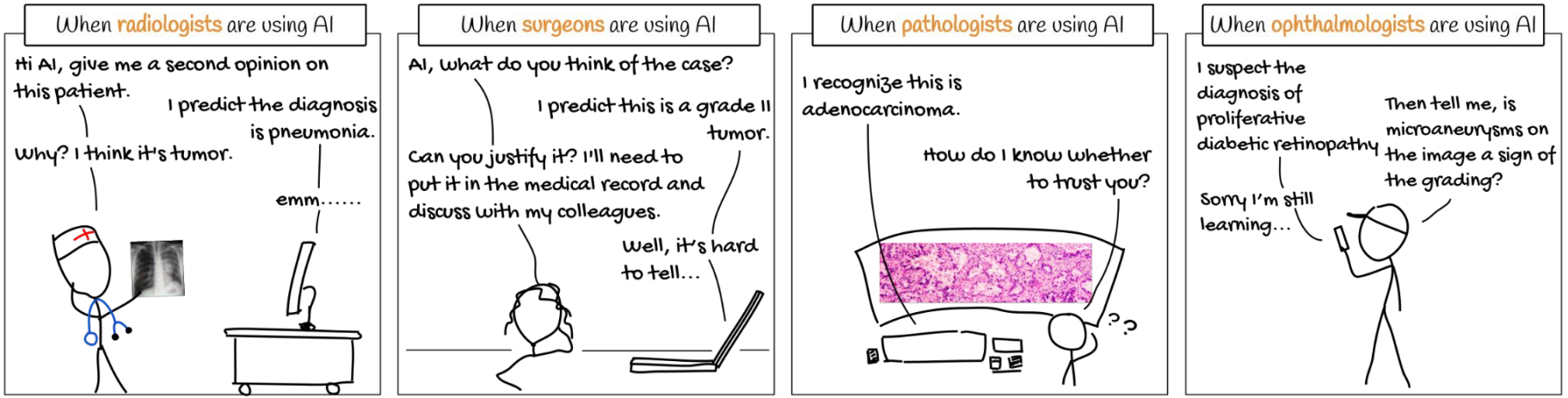
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

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?

Research paper

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

Weina Jin [a,*,1], Mostafa Fatehi [b,1], Ru Guo [b], Ghassan Hamarneh [a]

[a] School of Computing Science, Simon Fraser University, Burnaby, Canada
[b] Division of Neurosurgery, The University of British Columbia, Vancouver, Canada

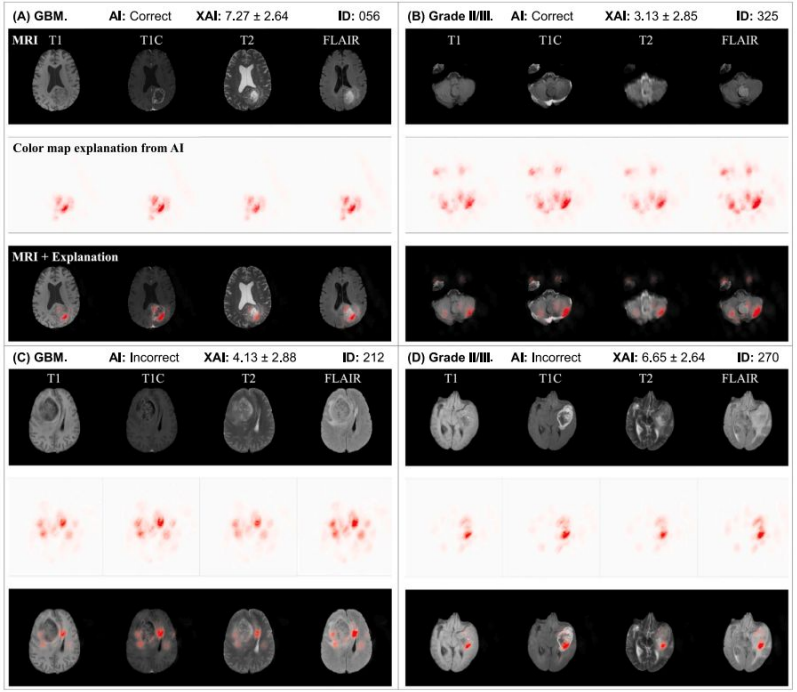
ARTICLE INFO

Dataset link: https://github.com/weinajin/multimodal_explanation

Keywords:
Artificial intelligence
Neuro-imaging
Neurosurgery
Explainable artificial intelligence
Clinical study
Human-centered artificial intelligence

ABSTRACT

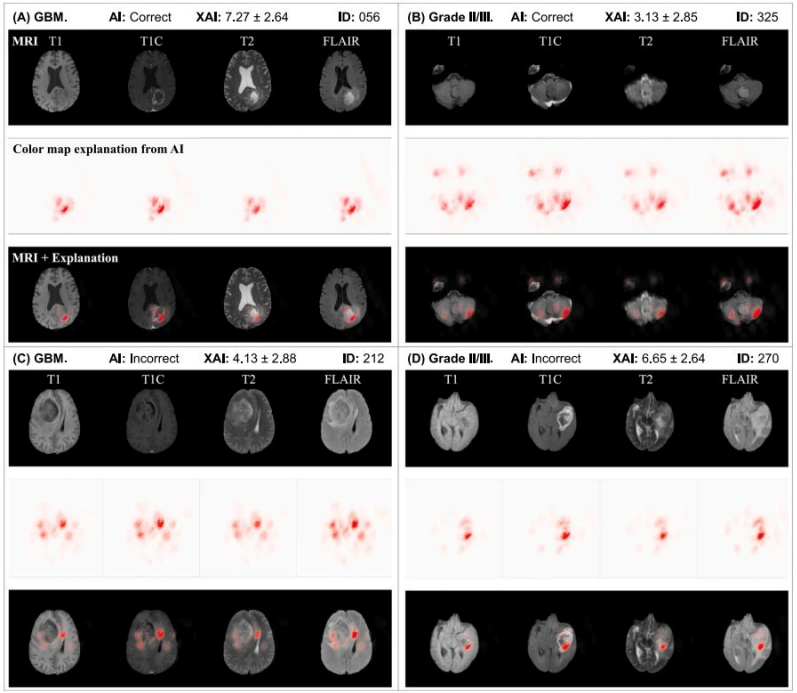
Clinical evaluation evidence and model explainability are key gatekeepers to ensure the safe, accountable, and effective use of artificial intelligence (AI) in clinical settings. We conducted a clinical user-centered evaluation with 35 neurosurgeons to assess the utility of AI assistance and its explanation on the glioma grading task. Each participant read 25 brain MRI scans of patients with gliomas, and gave their judgment on the glioma grading without and with the assistance of AI prediction and explanation. The AI model was trained on the BraTS dataset with 88.0% accuracy. The AI explanation was generated using the explainable AI algorithm of SmoothGrad, which was selected from 16 algorithms based on the criterion of being truthful to the AI decision process. Results showed that compared to the average accuracy of $82.5 \pm 8.7\%$ when physicians performed the task alone, physicians' task performance increased to $87.7 \pm 7.3\%$ with statistical significance (p-value = 0.002) when assisted by AI prediction, and remained at almost the same level of $88.5 \pm 7.0\%$ (p-value = 0.35) with the additional assistance of AI explanation. Based on quantitative and qualitative results, the observed improvement in physicians' task performance assisted by AI prediction was mainly because physicians' decision patterns converged to be similar to AI, as physicians only switched their decisions when disagreeing with AI. The insignificant change in physicians' performance with the additional assistance of AI explanation was because the AI explanations did not provide explicit reasons, contexts, or descriptions of clinical features to help doctors discern potentially incorrect AI predictions. The evaluation showed the clinical utility of AI to assist physicians on the glioma grading task, and identified the limitations and clinical usage gaps of existing explainable AI techniques for future improvement.



Weind Jin, Mostafa Fatehi (co-first authors), et al. Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. Artificial Intelligence in Medicine. 2024

What does “clinical user-centered” really mean?

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.” (N1)



Weind Jin, Mostafa Fatehi (co-first authors), et al. Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. Artificial Intelligence 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]

Input

AI model

AI's suggestion:

A fly

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

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

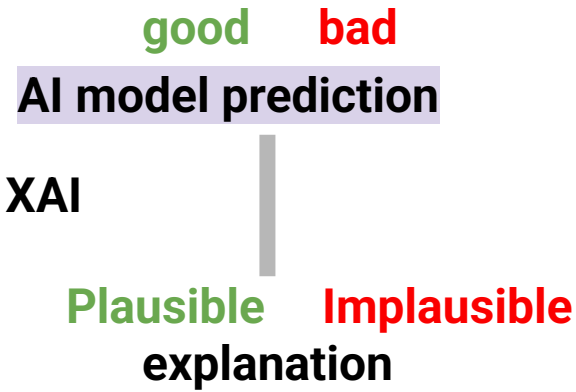
=

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

Why is plausibility prevalently used in XAI evaluation?



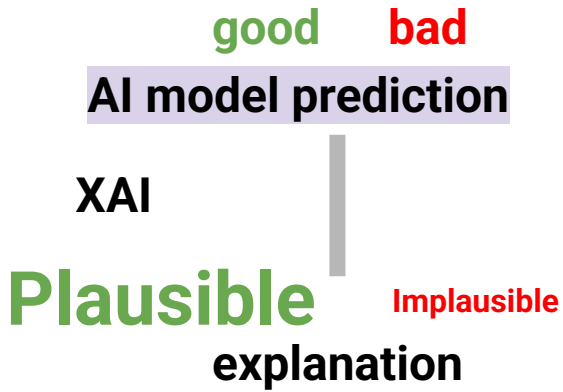
Plausibility measures how well an *AI explanation* aligns with *human explanations*.

⇨ 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

Why is plausibility prevalently used in XAI evaluation?



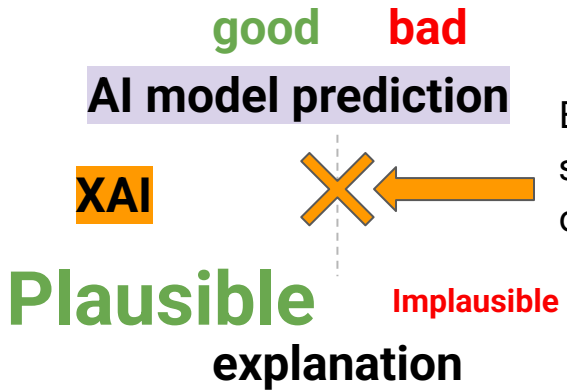
Plausibility measures how well an *AI explanation* aligns with *human explanations*.

⇨ 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

Why is plausibility prevalently used in XAI evaluation?



But this is not what XAI is supposed to do, and can destroy XAI's original function

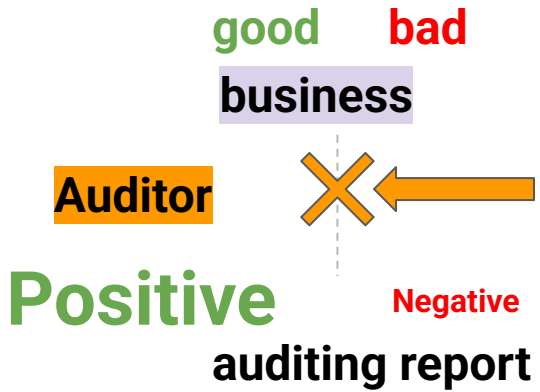
Plausibility measures how well an *AI explanation* aligns with *human explanations*.

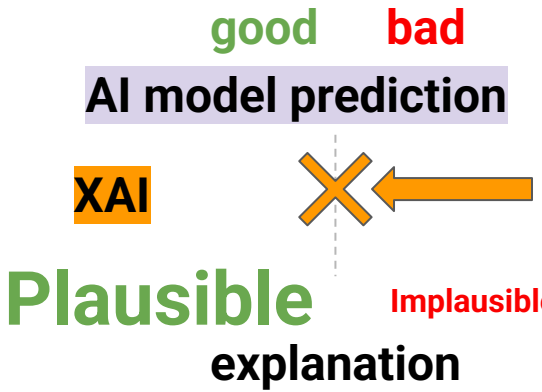
Optimize/evaluate XAI using plausibility

We want more plausible explanations

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

Why is plausibility prevalently used in XAI evaluation?





But this is not what XAI is supposed to do, and can destroy XAI's original function

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

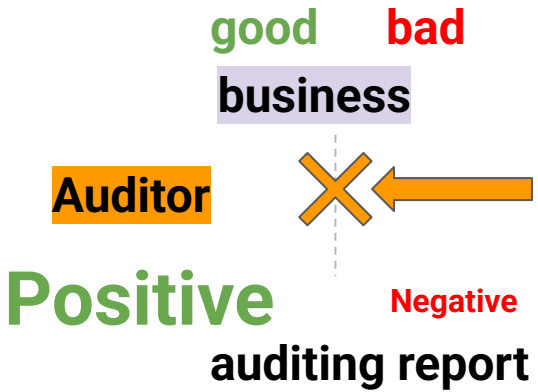
Optimize/evaluate XAI using plausibility

We want more plausible explanations

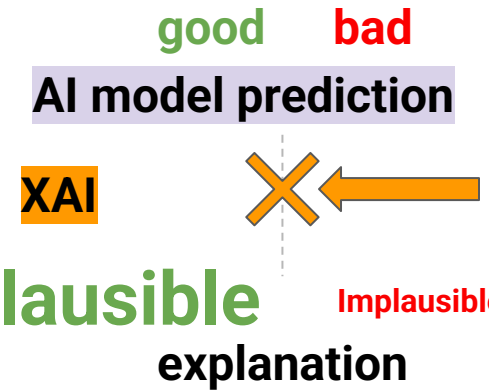
[1] Why is plausibility surprisingly problematic as an XAI criterion? Weina Jin, Xiaoxiao Li, Ghassan Hamarneh. arXiv:2303.17707

Why is plausibility prevalently used in XAI evaluation?

Incentivize (bribe) auditors to generate more positive reports



Incentivize (bribe) XAI algorithms to generate more plausible explanations



But this is not what XAI is supposed to do, and can destroy XAI's original function

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

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

- Explainability is one of the AI ethics principles that enables the other ethics principles.
- Evaluation methods guide the development of a field.
- If this sentinel field of XAI has this concerning phenomenon, other technical fields may be more concerning.
- Are we really on the right track?
  - I provide a counterexample
  - Paths that have the least resistance do not necessarily mean they are the right ones.
  - Technical practices are conducted in a sociotechnical system.
  - These paths can be shaped by hidden power and disguised as natural paths. —> manipulation!

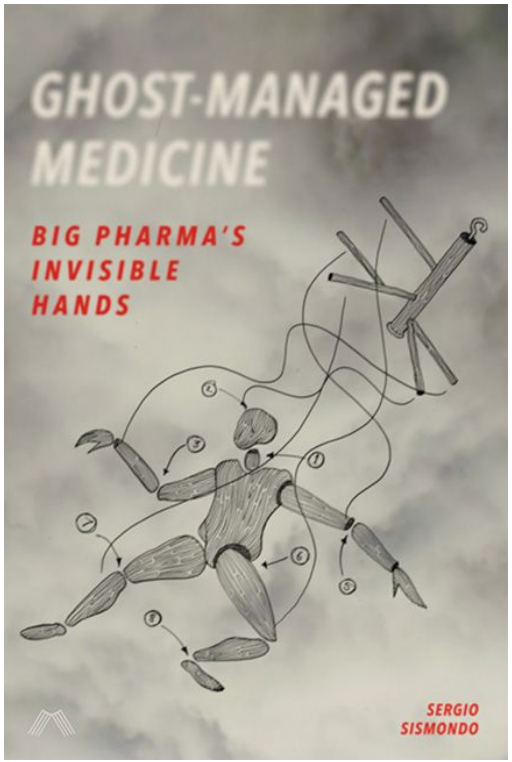
Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
   a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Are we really on the right track?

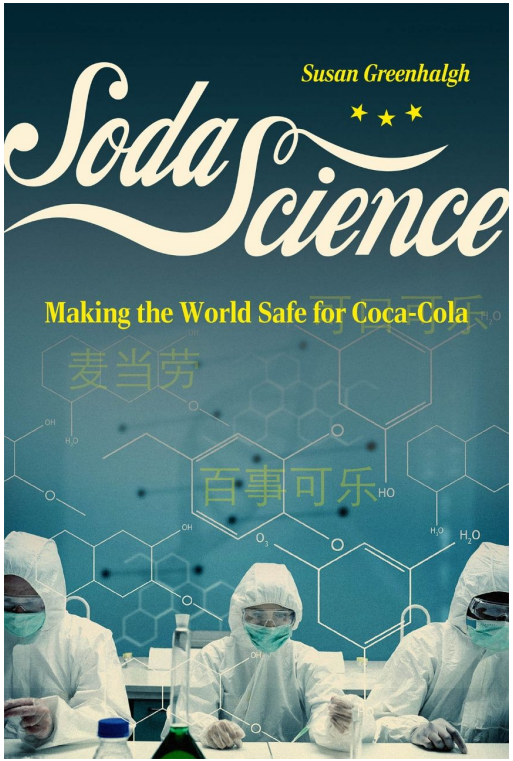
Or is it an illusion created to keep us manipulable?



"**Pharmaceutical companies** sustain large networks to gather, create, control and disseminate information. They provide the pathways that carry this information, and the energy that makes it move. Through bottlenecks and around curves, knowledge is created, given shape by the channels it navigates. Pharma companies create medical knowledge and move it to where it is most useful; much of it is **perfectly ordinary knowledge** that happens to support their marketing goals. But because of the companies' resources, their interests and their levels of control, **they become key shapers of almost all medical terrains**."

Are we really on the right track?

Or is it an illusion created to keep us manipulable?



Anthropologist Susan Greenhalgh tells the story of how **industry leader Coca-Cola** collaborated with academia to conduct **real scientific research** that advocated exercise, not calorie restraint, as the priority solution for obesity.

This **distorted research agenda** influenced public health policies on obesity and public understanding on diet and lifestyle in favor of the needs and profits of soda industry.

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
    a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

Critical technical practice

Assumptions

critically inspect assumptions and justifications in common technical practices

We cannot make such a certain claim without critical reflections and examinations.

What if the track/root is wrong? Any post-hoc fix won't help.

Is it really natural and inevitable? Are there any alternatives?

Potential clinical utilities of AI explanations: the clinical users' aspect

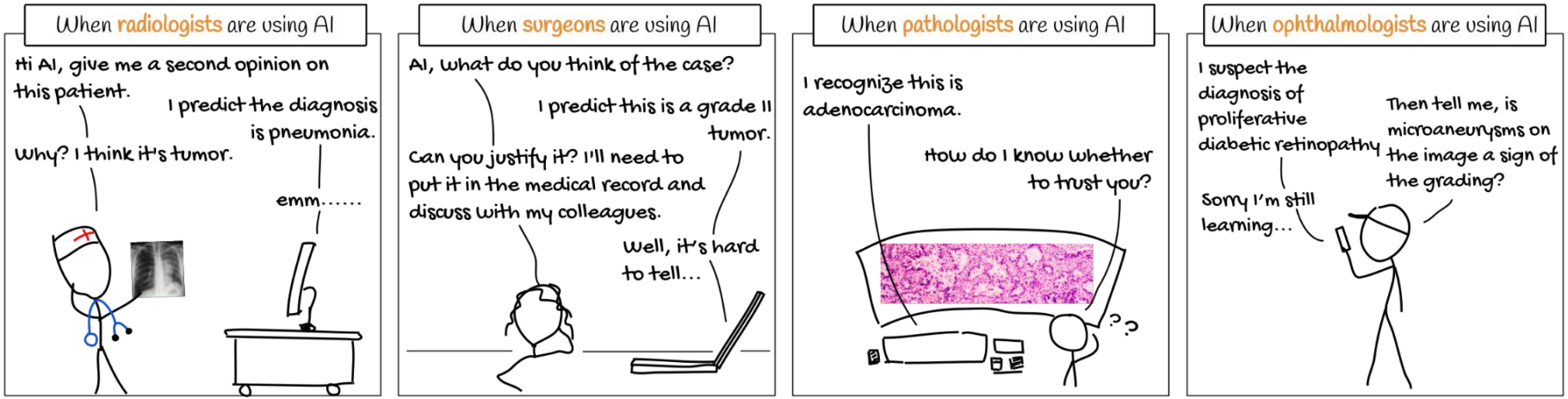
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

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

Potential clinical utilities of AI explanations: the clinical users' aspect

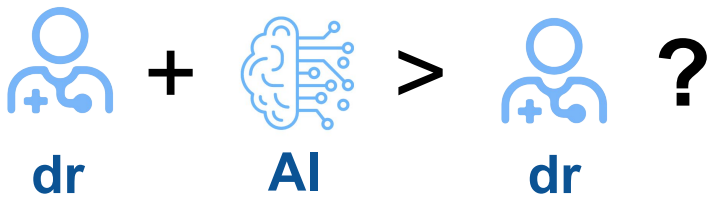
Explainable AI (XAI): To explain AI predictions in human-understandable ways [2]

Resolve clinical users' disagreement with AI [1]	Verify AI's suggestion [1]	*Calibrate* clinical users' trust [1]	Learning & medical knowledge discovery [1]
informed decision	evidence-based medicine		

Research question:

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

Hypothesis:



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

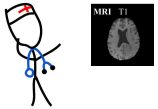
Study design

Each doctor underwent **three conditions**

Measurement: **task performance**

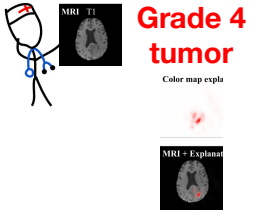
1. **DR**

Doctor only



2. **DR+AI**

Doctor assisted by AI **prediction**



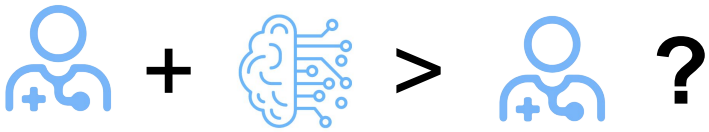
3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**

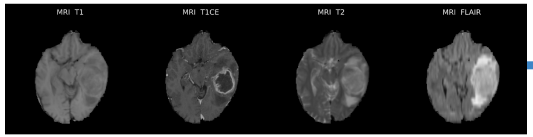
Research question:

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

Hypothesis:



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



Input

AI model

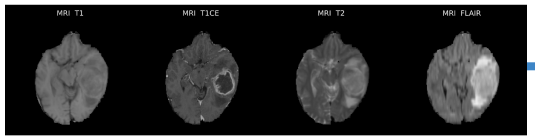
AI's suggestion:
High-grade brain tumor

Output

[1] Smoothgrad: removing noise by adding noise, 2017. Arxiv: 1706.03825

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

AI assistance: a suggestion + an explanation



Input

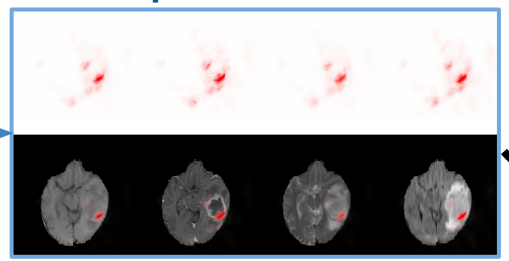
AI model

3D convolutional neural network

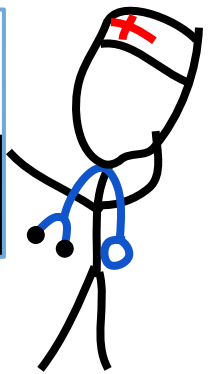
AI's suggestion:
High-grade brain tumor

Accuracy: 88%

AI's explanation:



SmoothGrad XAI algorithm [1]

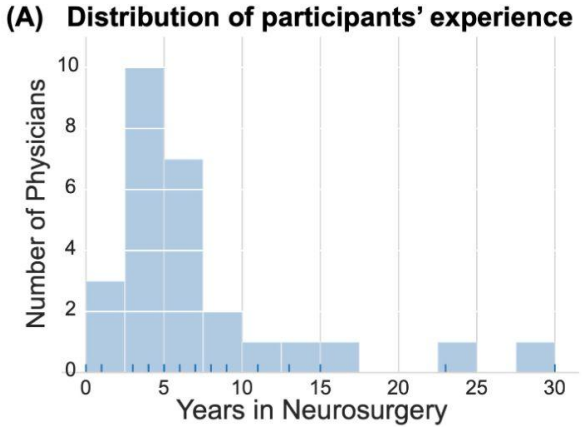


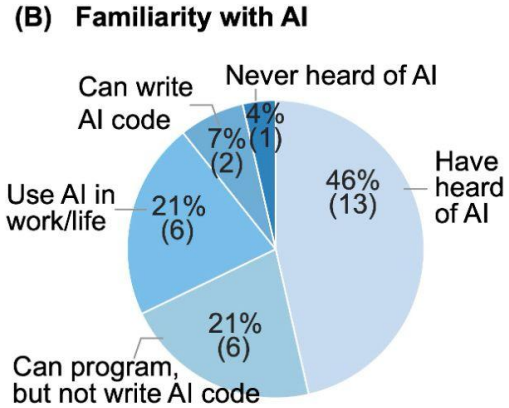
[1] Smoothgrad: removing noise by adding noise, 2017. Arxiv: 1706.03825

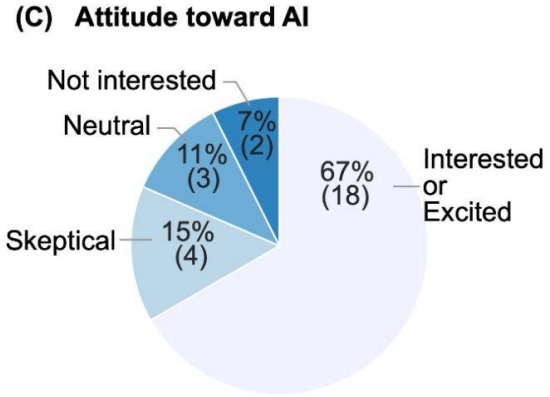
A user study on the clinical utility of AI and its explanation in a brain tumor grading task

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





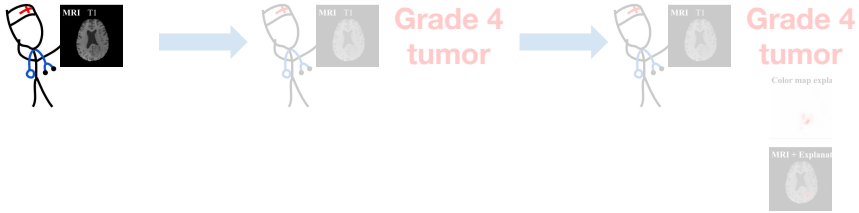


Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

Study procedure

35 neurosurgeons, each read 25 brain images

Drs gave judgment on three conditions:



1. **DR**

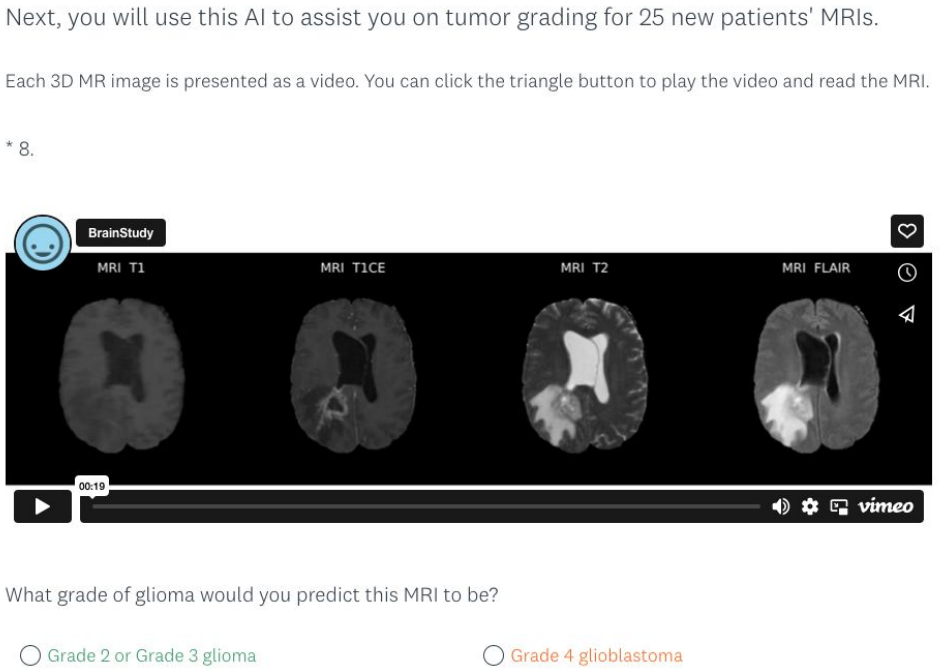
Doctor only

2. **DR+AI**

Doctor assisted by AI **prediction**

3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**

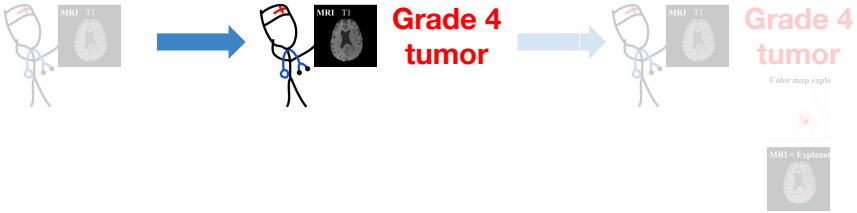


Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

Study procedure

35 neurosurgeons, each read 25 brain images

Drs gave judgment on three conditions:



1. DR

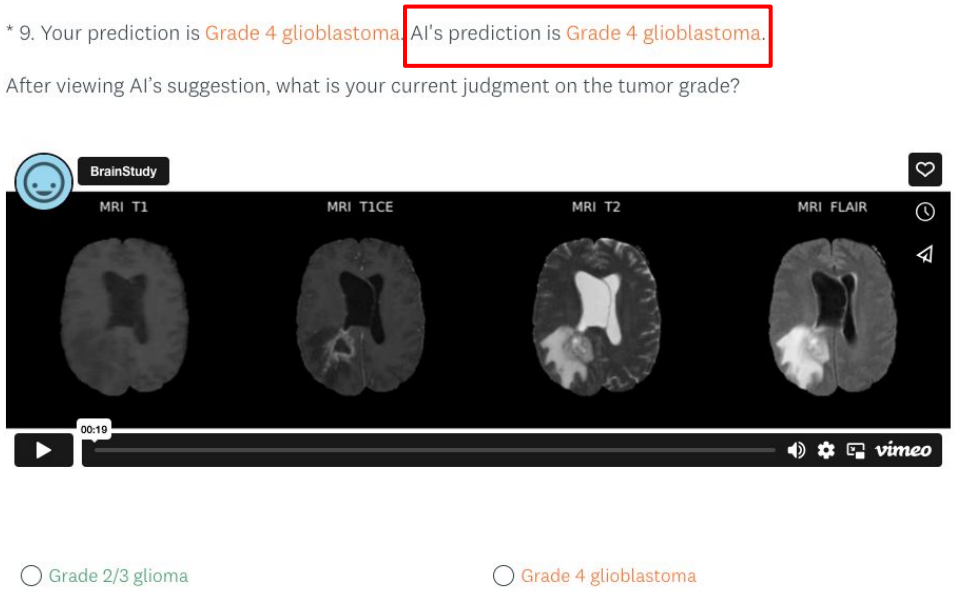
Doctor only

2. **DR+AI**

Doctor assisted by AI **prediction**

3. DR+XAI

Doctor assisted by AI prediction & explanation

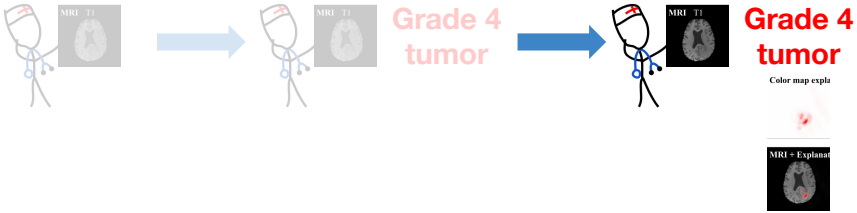


Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

Study procedure

35 neurosurgeons, each read 25 brain images

Drs gave judgment on three conditions:



1. **DR**

Doctor only

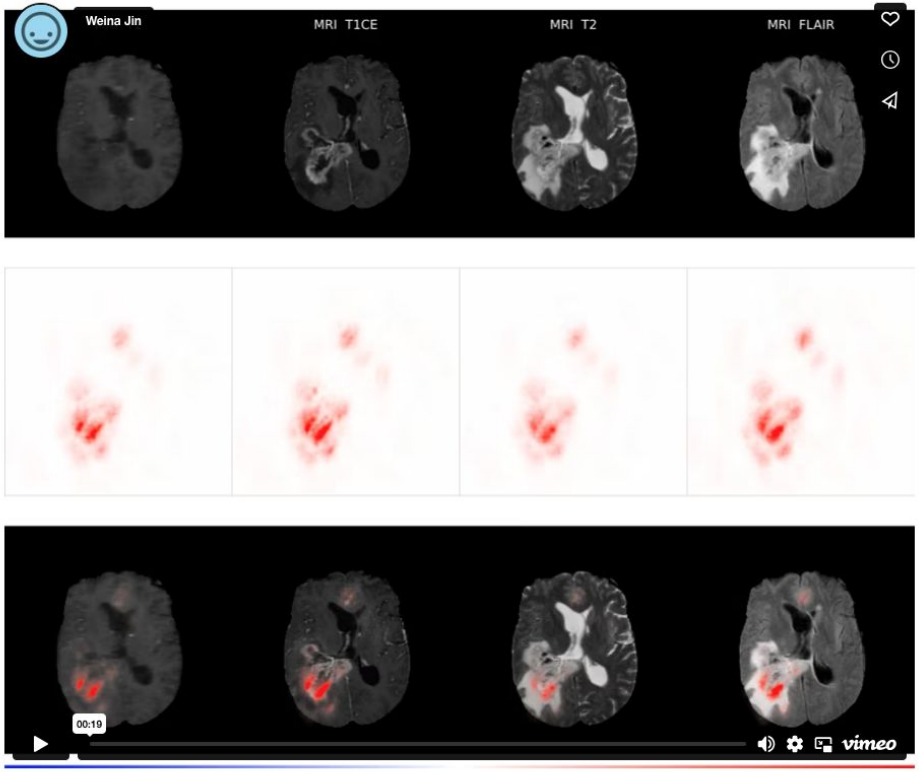
2. **DR+AI**

Doctor assisted by AI **prediction**

3. **DR+XAI**

Doctor assisted by AI prediction & **explanation**

* 11. Please click the triangle "play" button to play the video and read the color map explanation.

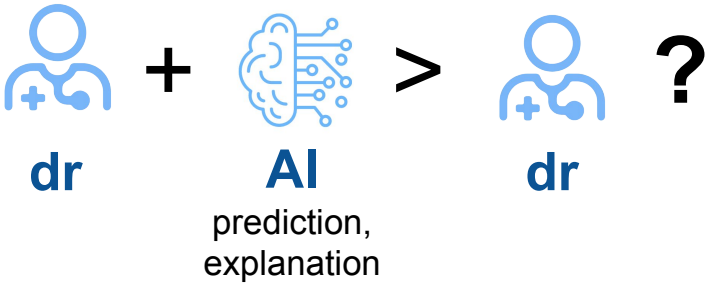


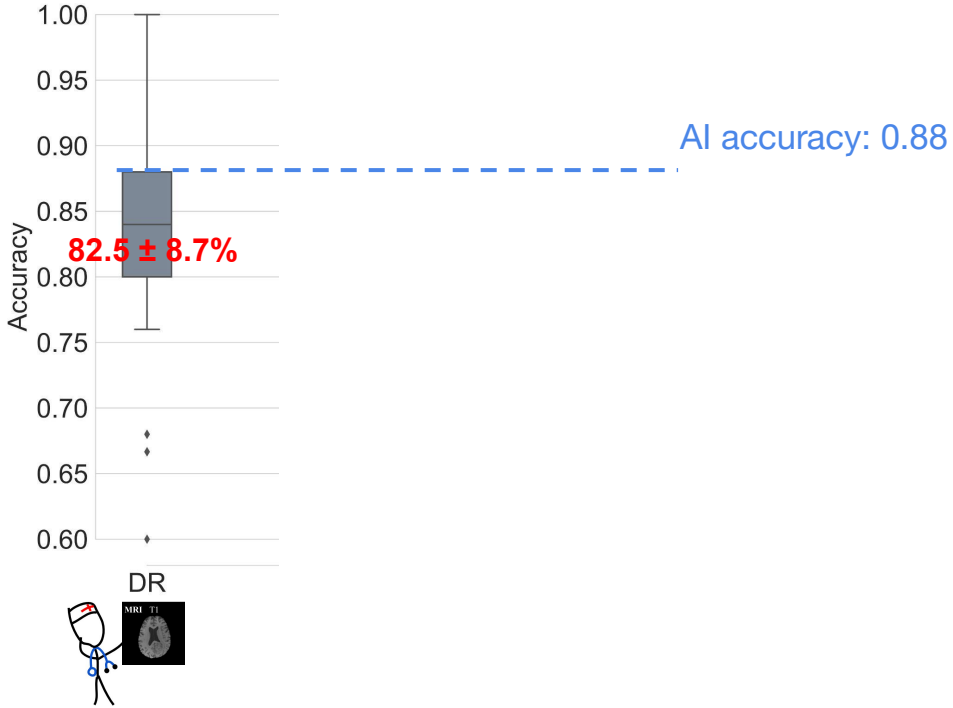
After viewing AI's explanation, what is your final judgment on the tumor grade?

◯ Grade 2/3 glioma ◯ Grade 4 glioblastoma

Result

Is doctor+AI better than doctor alone?

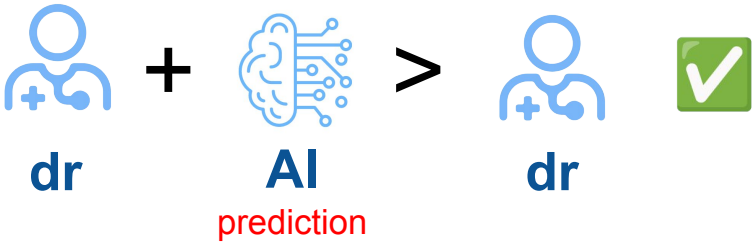


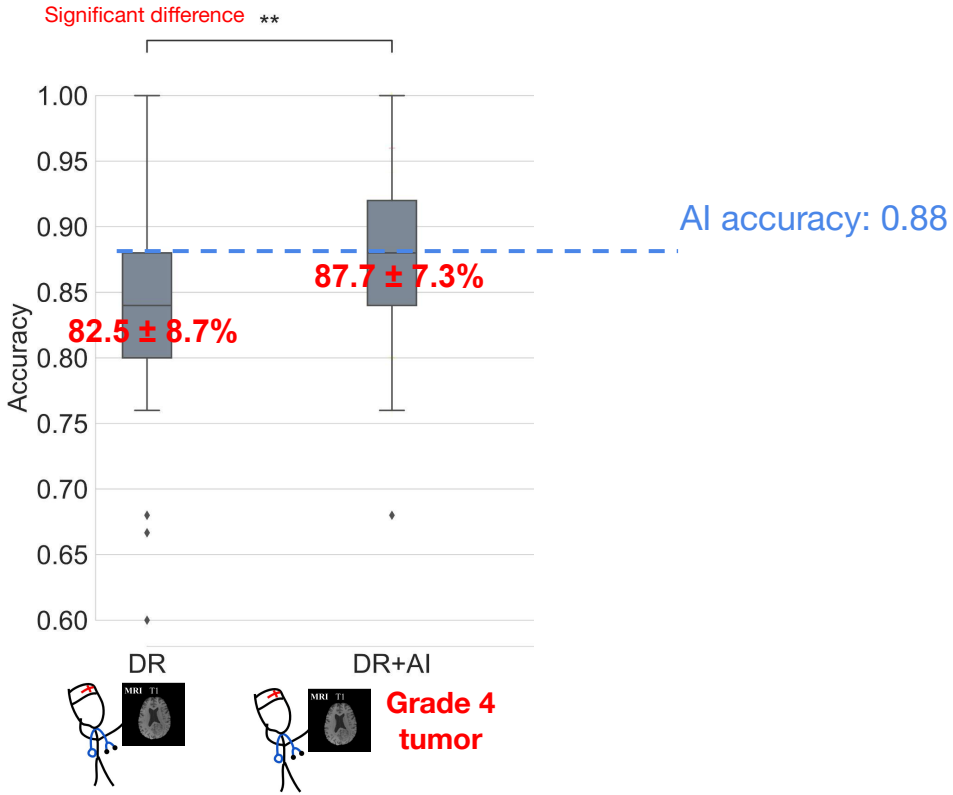


Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task. W Jin, M Fatehi, R Guo, G Hamarneh. AI in Medicine. 2024

Result

Is doctor+AI better than doctor alone?





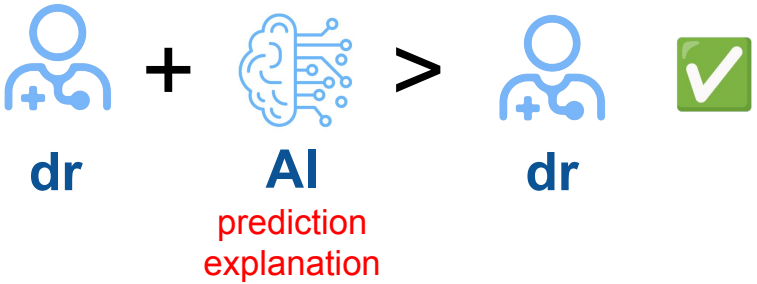
In an experimental setting:

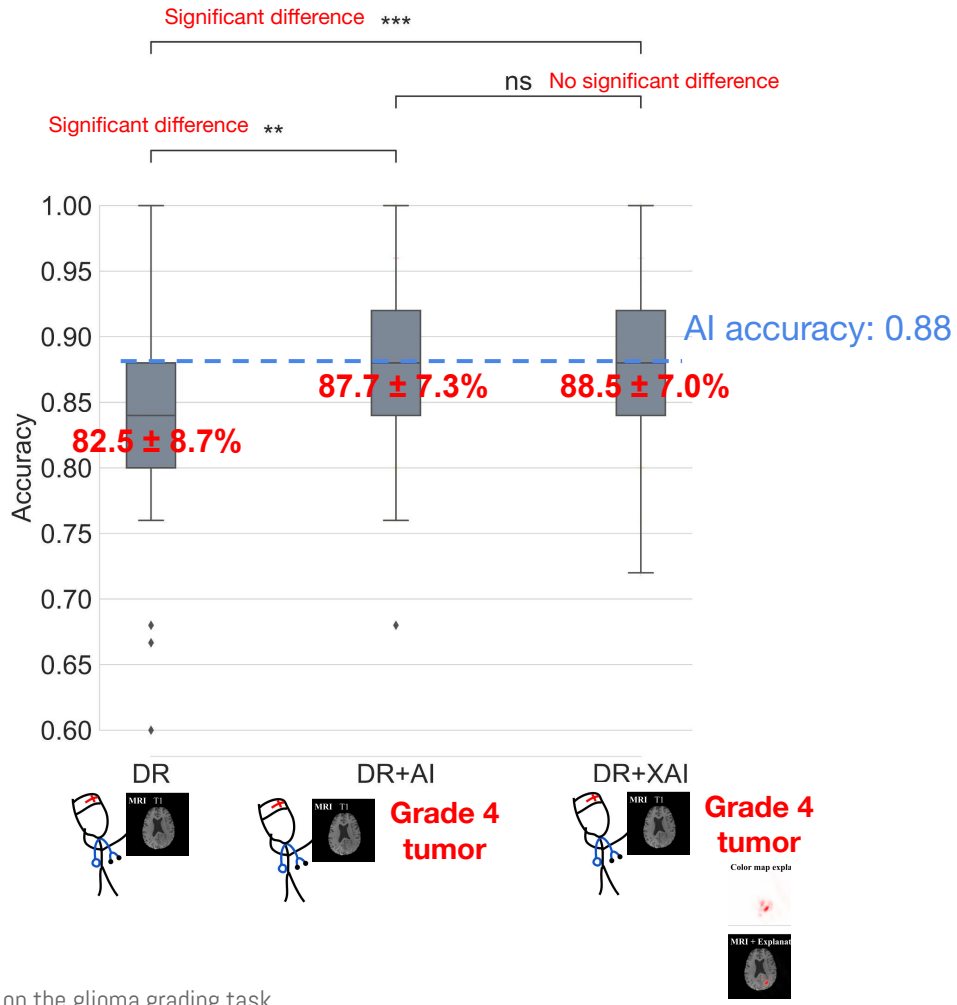
- AI prediction (**DR+AI**) is helpful to improve doctors' task performance

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

Result

Is doctor+AI better than doctor alone?





In an experimental setting:

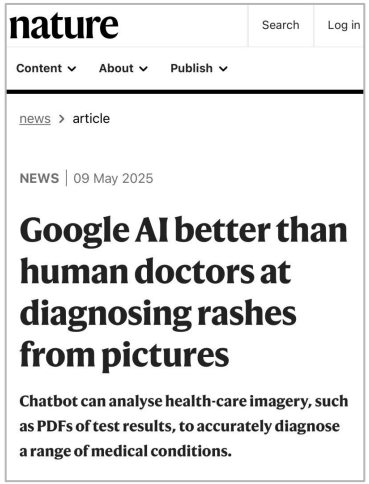
- AI prediction (**DR+AI**) is helpful to improve doctors' task performance
- AI explanation (**DR+XAI**) does not show additional help in improving doctors' performance

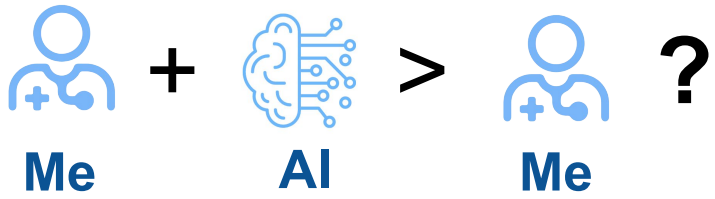
Evaluating the clinical utility of artificial intelligence assistance and its explanation on the glioma grading task.

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

Similar findings and claims on

AI outperforms humans & human performance improved with AI assistance
in a, b, c, … tasks





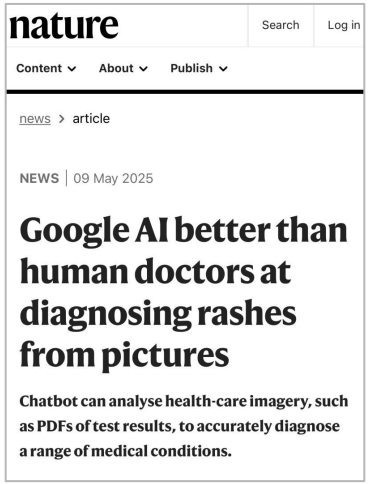
prediction,explanation

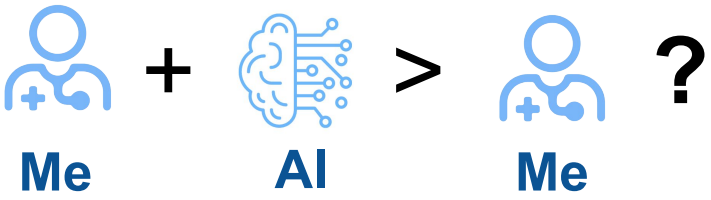
This study is relevant to similar issues when using AI to assist users

Similar findings and claims on

AI outperforms humans & human performance improved with AI assistance in a, b, c, … tasks

AI explanation/other human-centered properties fail in x, y, z, … tasks



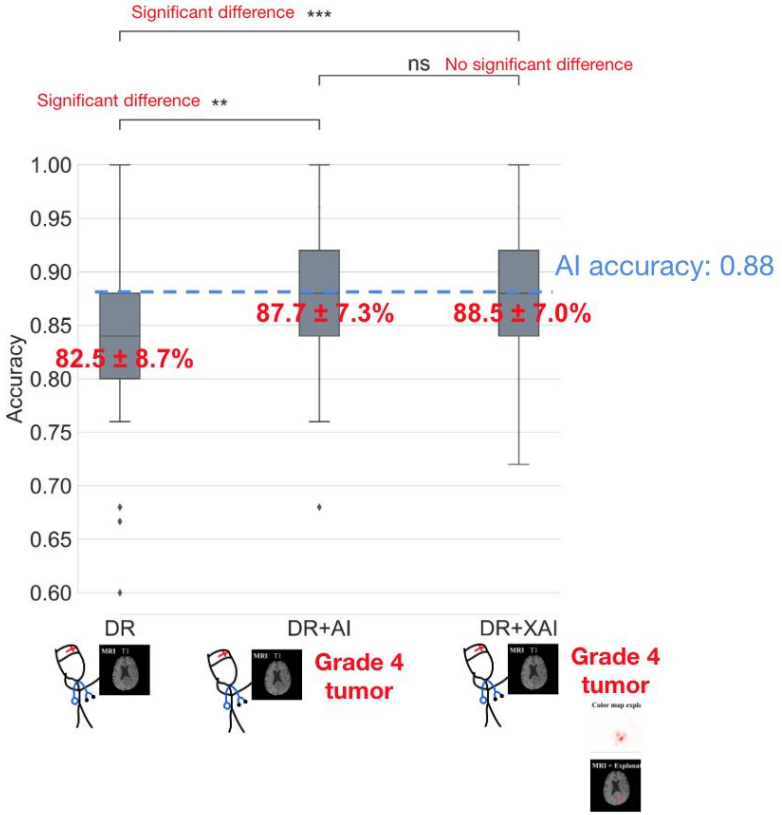


prediction,explanation

The same evidence can be interpreted differently to support opposite AI development goals

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

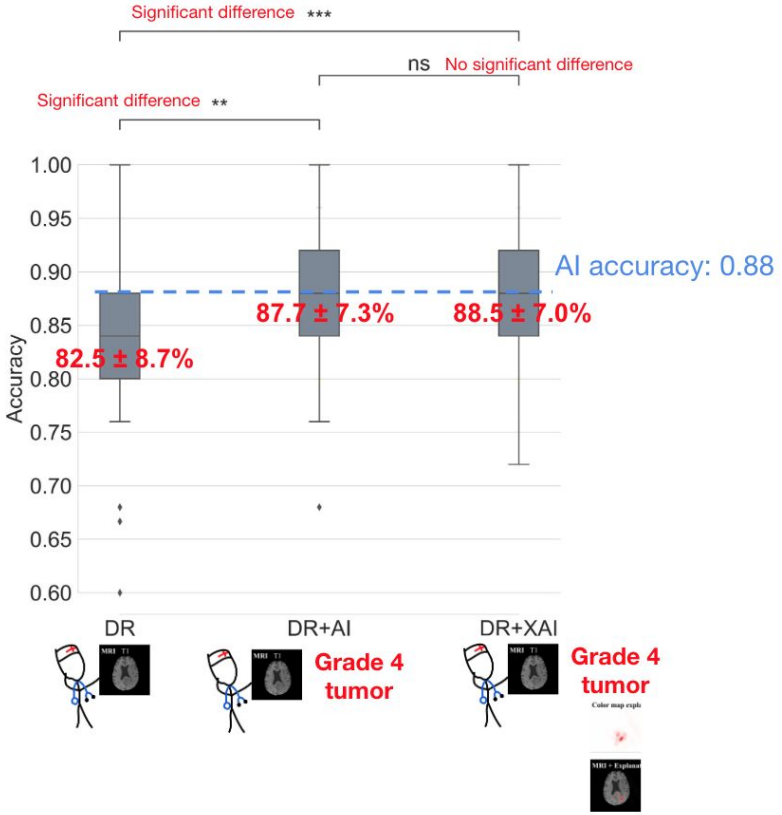
The same evidence can be interpreted differently to support opposite AI development goals

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

The same evidence can be interpreted differently to support opposite AI development goals

AI optimism → Develop AI to **compete** with humans, XAI **can be traded off** with high performance

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, XAI **cannot be traded off**

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**.

Technical practice

Assumptions

1. The technical development and practice are generally on the right track.
   a. Meaning, their limitations, mistakes, and errors can be fixed by continuous technical innovations & improvements.
2. The technical agenda is a natural process that reflects a scientific and technological development trajectory.

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

XAI does not trade off with performance

As long as the XAI can reliably indicate for humans **when to rely on AI, and when not to**, the human-AI team can achieve complementary performance that is better than either human or AI alone [1].

THEOREM 2 (CONDITIONS FOR XAI COMPLEMENTARITY). *Let h, m, and t be the accuracies of the human, AI, and human-AI team, respectively; $f(P_i)$ be the probability from a Bernoulli distribution of human acceptance of an AI suggestion for an instance $x_i \in \mathcal{D}$, where $f(.)$ is a monotonically non-decreasing function of the explanation plausibility P_i; and P_i^r and P_i^w be the plausibility values of an AI explanation when an instance x_i is predicted correctly or incorrectly, then:*

Complementary human-AI accuracy can be achieved, i.e., $t > \max(h, m)$, when

$$\begin{cases} h \geq m \quad and \quad \mathbb{E}[f^r] > \frac{h(1-m)}{m(1-h)}\mathbb{E}[f^w]; & or, \\ m > h \quad and \quad \mathbb{E}[f^r] > \frac{h(1-m)}{m(1-h)}\mathbb{E}[f^w] + \frac{m-h}{m(1-h)} \end{cases}$$

where $\mathbb{E}[f^r]$ and $\mathbb{E}[f^w]$ are the expectations of $f(P_i^r)$ and $f(P_i^w)$ over the dataset $\mathcal{D}$, indicating among the correctly or incorrectly predicted instances of AI, how many are accepted by human.

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

XAI does not trade off with performance

Develop AI to **compete** with humans, XAI **can be traded off** with high performance

As long as the XAI can reliably indicate for humans **when to rely on AI, and when not to**, the human-AI team can achieve complementary performance that is better than either human or AI alone [1].

An AI objective that is totally different from the current mainstream AI paradigm

Develop AI to **collaborate** with humans, XAI **cannot be traded off**

THEOREM 2 (CONDITIONS FOR XAI COMPLEMENTARITY). *Let h, m, and t be the accuracies of the human, AI, and human-AI team, respectively; $f(P_i)$ be the probability from a Bernoulli distribution of human acceptance of an AI suggestion for an instance $x_i \in \mathcal{D}$, where $f(.)$ is a monotonically non-decreasing function of the explanation plausibility P_i; and P_i^r and P_i^w be the plausibility values of an AI explanation when an instance x_i is predicted correctly or incorrectly, then:*

Complementary human-AI accuracy can be achieved, i.e., $t > \max(h, m)$, when

$$\begin{cases} h \geq m & \text{and} \quad \mathbb{E}[f^r] > \frac{h(1-m)}{m(1-h)}\mathbb{E}[f^w]; & \text{or,} \\ m > h & \text{and} \quad \mathbb{E}[f^r] > \frac{h(1-m)}{m(1-h)}\mathbb{E}[f^w] + \frac{m-h}{m(1-h)} \end{cases}$$

where $\mathbb{E}[f^r]$ and $\mathbb{E}[f^w]$ are the expectations of $f(P_i^r)$ and $f(P_i^w)$ over the dataset $\mathcal{D}$, indicating among the correctly or incorrectly predicted instances of AI, how many are accepted by human.

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

Why is XAI (or other human-centered properties) **often framed as a trade-off with performance?**

Performance/accuracy/efficiency is the dominant value in AI [1].

Our study shows the value of performance has **an invasive nature** that can gradually **exclude** other ethical values, just like invasive species [2].

trade-off

[1] The Values Encoded in Machine Learning Research. Abeba Birhane, et al. 2022.
[2] 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

Good intentions do not necessarily lead to **good outcomes**

Human-centered technology

Clinical User-Centred Explainable AI for Medical Image Analysis

Explainability AI for Healthcare

Algorithmic Accountability

Ethical AI

Technology for Good

Diversity

Equity

Fairness

How can we make our ethical and responsible pursuit in technical development *effective* and *robust* ?

Human-centered technology

AI for Social Good

AI for science

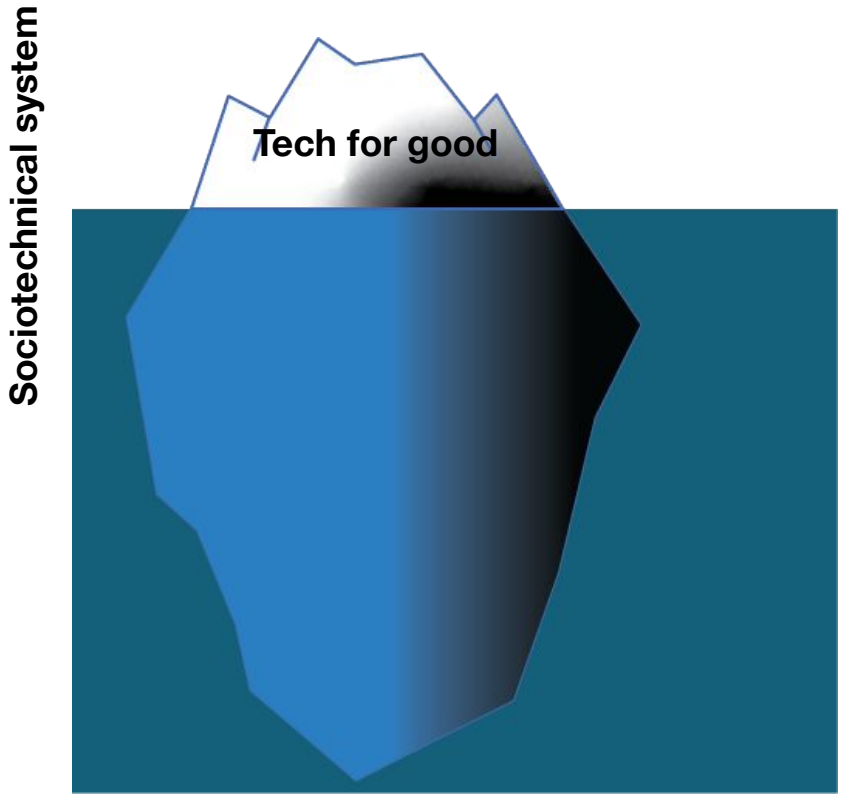
AI Justice

Explainability

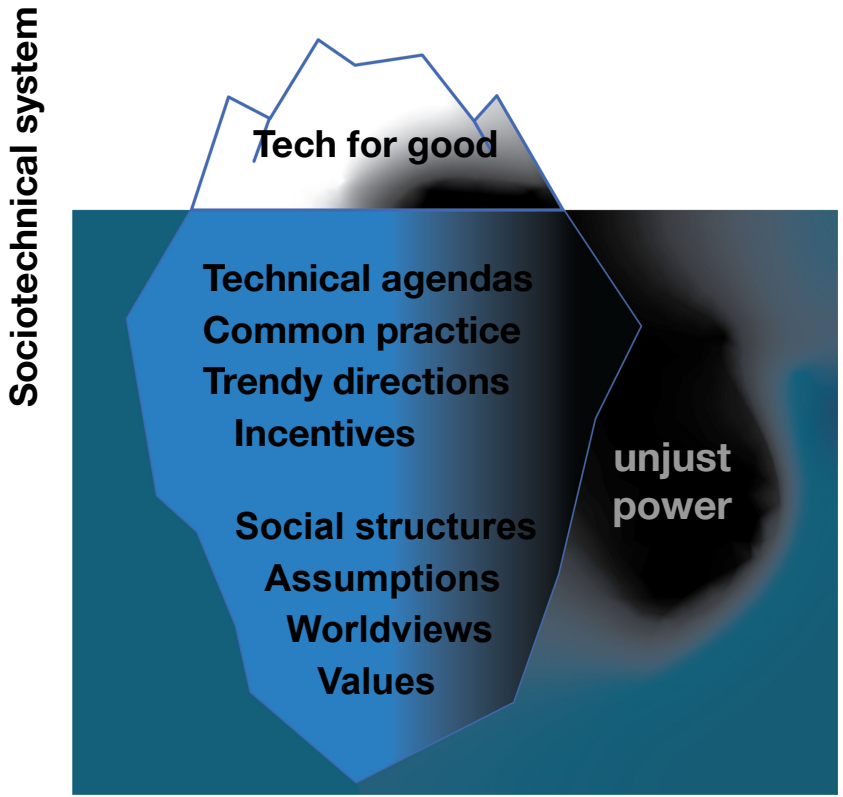
AI for Healthcare

Responsible Conduct


Tech for good



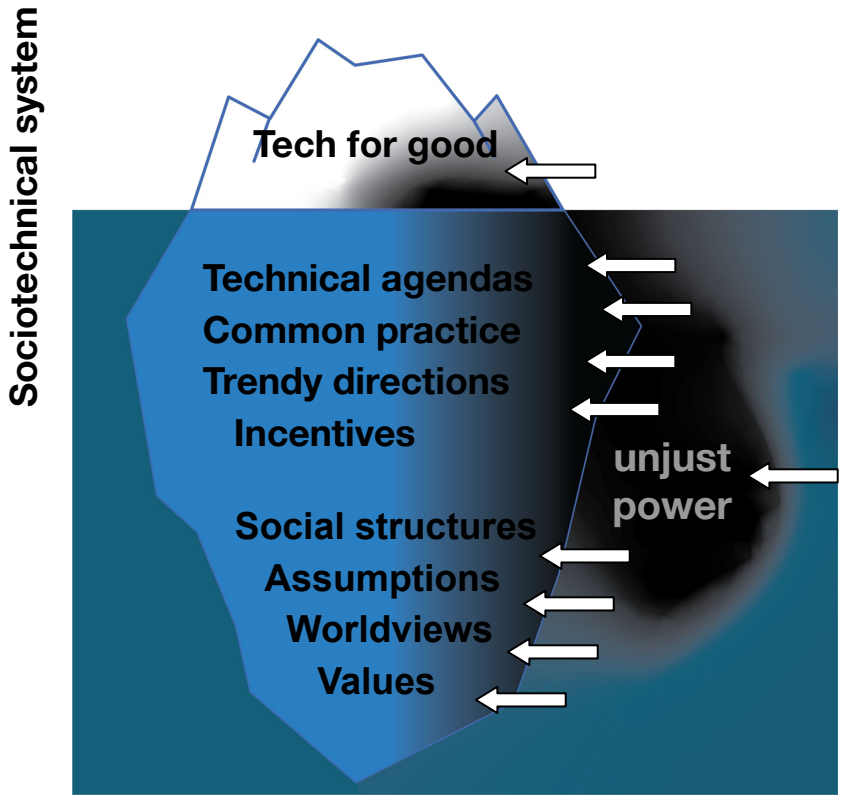
Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice.
Weina Jin, Elise Li Zheng, Ghassan Hamarneh. 2025. arXiv:2510.10588



Root problem

The unjust power, which has penetrated into all the components in the sociotechnical system that underpins a technology [2]

[1] Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. Philip E. Agr. 1997
[2] 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



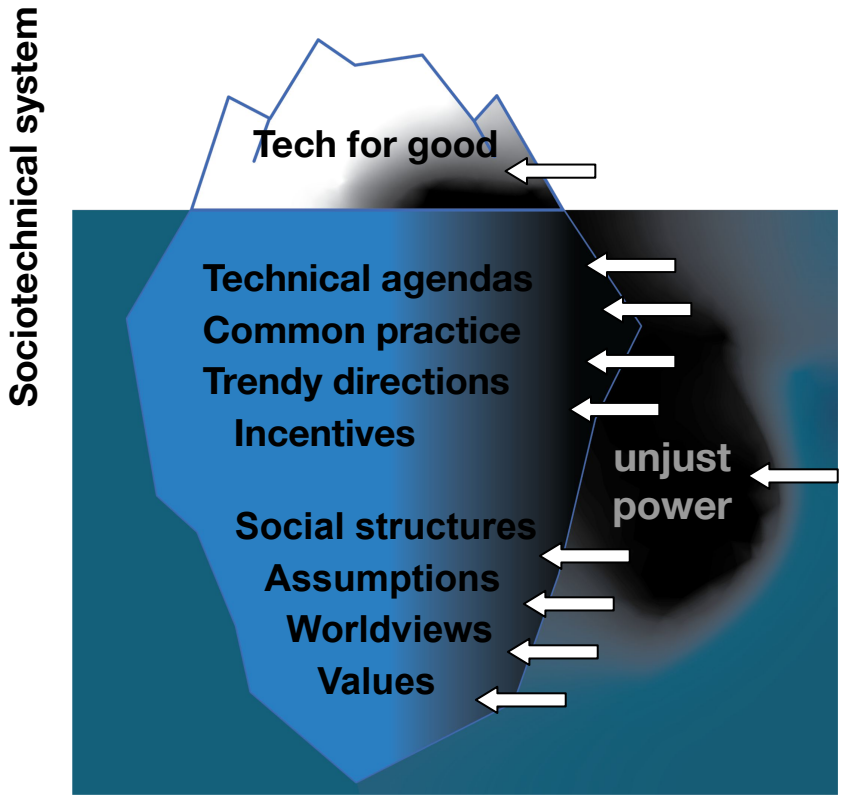
Strengthen the "immune system" in the technical ecosystem

systematically tackle the root cause and all the key points in the power mechanism

Root problem

The unjust power, which has penetrated into all the components in the sociotechnical system that underpins a technology [2]

[1] Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. Philip E. Agr. 1997
[2] 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



Critical Technical Practice [1]

Strengthen the "immune system" in the technical ecosystem

systematically tackle the root cause and all the key points in the power mechanism

Root problem

The unjust power, which has penetrated into all the components in the sociotechnical system that underpins a technology [2]

[1] Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. Philip E. Agr. 1997
[2] 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

Algorithmic Accountability

Ethical AI

Technology for Good

Diversity

Equity

Fairness

Critical Technical Practice

Method and demonstration of intervention

AI Justice

Human-centered technology

AI for Social Good

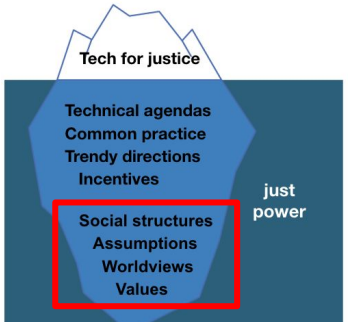
AI for Healthcare

Explainability

AI for science

Responsible Conduct

A new imagination of AI: AI for just work



Mainstream imagination of AI & its questionable assumptions (Sec. 2)

S1. *AI can be capable of solving many hard problems*

It surreptitiously substitutes model for the real, and ignores AI's fundamental dependences on human experience and natural resources.

S2. *AI replacement can free humans*

It forgets AI's fundamental dependence on labor, devalues workers and degrades work to make AI appear "automatic."

S3. *AI can improve productivity*

It is more likely to lead to inequality, as it inherently lacks power check mechanism to ensure shared prosperity.

Demonstrating our process of constructing a new imagination: AI for just work (Sec. 4)

Baking ethics into the foundational assumptions (Sec. 4.1)	Deducing AI properties from assumptions (Sec. 4.2)	Applying the new imagination in AI technical development (Sec. 4.3)
A1. *AI has intrinsic limits and fundamental dependences*	P1. *AI is grounded in real-world problems*	Propose ethical criteria and evaluation paradigm to ground AI in real-world problems
A2. *Worker and work are not devalued or degraded*	P2. *Thorough limitation analysis of AI is a default*	Develop three aspects of limitation analysis to assess and declare AI's limitations
A3. *Prioritize ethics over productivity*	P3. *AI and data collection are rejectable*	Develop a stakeholders' checklist to question the appropriateness of AI
	P4. *Choose to design AI that can increase workers' power*	

AI for Just Work: Constructing Diverse Imaginations of AI beyond "Replacing Humans"
Weina Jin, Nicholas Vincent, Ghassan Hamarneh

Critical Technical Practice

Demonstration in the medical image synthesis task

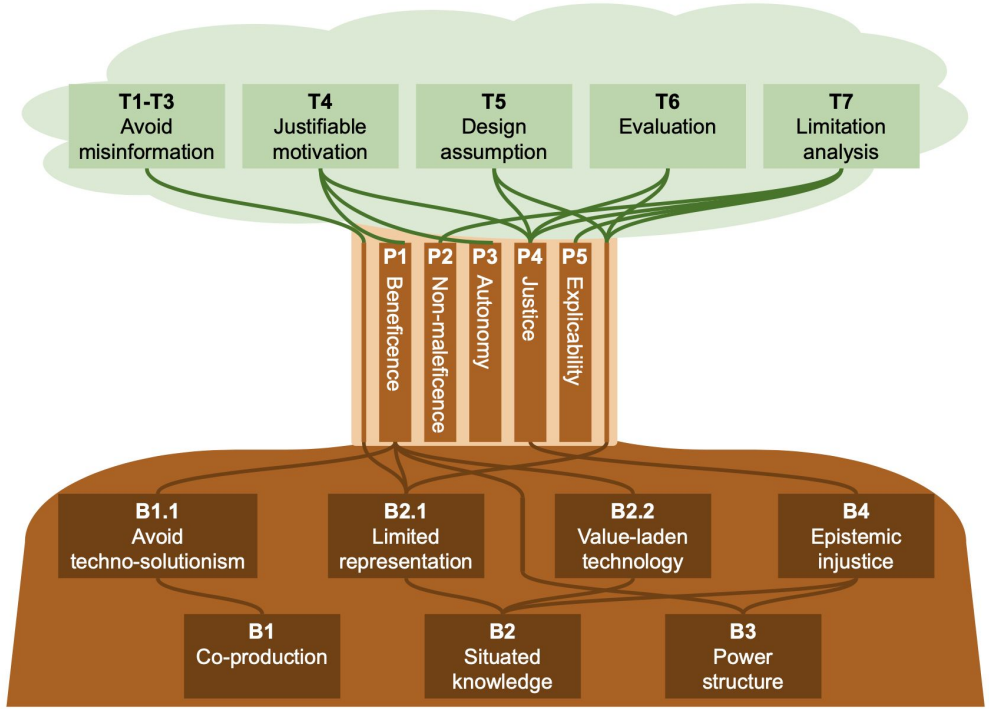
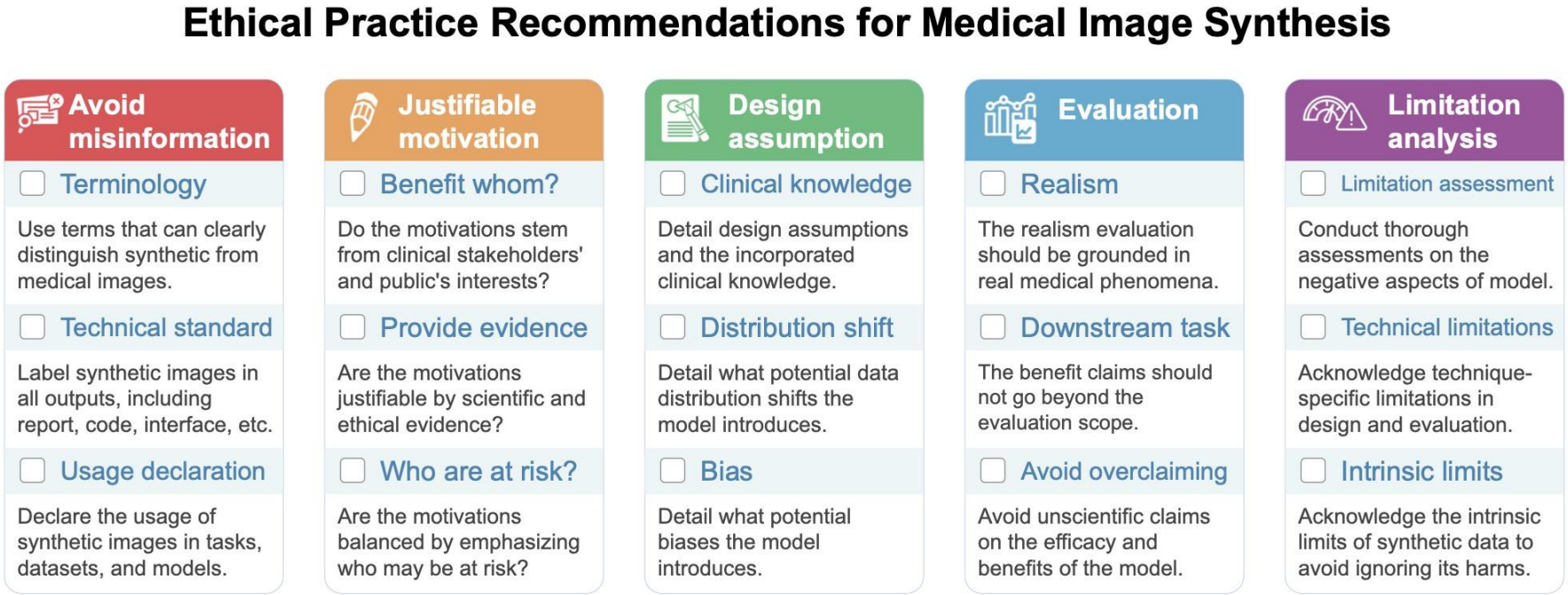


Figure 3: A tree structure visualizing the relationship of the background assumptions B1-B4 (roots) in Appendix A2, the ethical principles P1-P5 (trunk) in Appendix A3, and the Ethical Practice Recommendations T1-T7 (leaves) in Section 2. The roots and trunk of the ethical theories provide theoretical supports (visualized as lines) for the leaves of the Ethical Practice Recommendations.

Weina Jin, et al. Ethical Medical Image Synthesis. 2025. arXiv:2508.09293

Critical Technical Practice

Demonstration in the medical image synthesis task



Weina Jin, et al. Ethical Medical Image Synthesis. 2025. arXiv:2508.09293

Critical Technical Practice

Demonstration in the explainable AI task

Table 2: The mainstream values, worldview, and foundational assumptions in the technical community vs. ours.

	Mainstream worldview in the technical community	**Our worldview**
Core worldview	Theories and computational models can fully represent and transcend the real-world complex phenomena via reduction and abstraction.	The complex real-world phenomena are irreducible, incompressible, and cannot be fully represented by theories and computational models.
Core values	Performance, efficiency, innovation, competition, novelty, static, rule, prediction, certainty.	Ethics, justice, relationship, care, collaboration, dynamic, fluidity, vagueness, uncertainty.
Assumptions of technical development	Technical development happens in a highly abstracted setting, and is not influenced by social and political factors. Technology is apolitical and value-neutral.	Technology is value laden, shaped by social, political, and power structures, and in turn shapes society.
Assumptions of knowledge and reasoning	• There exists a perfect state of knowledge of all knowing.	• Knowing is always partial and situated. There does not exist a perfect state of knowledge or knower.
	• Humans are flawed reasoners, which can be transcended by some groups of humans and/or computational models.	• The imperfect state of knowledge or knower cannot be transcended. Knowing can always be improved by incorporating more perspectives from people who are socially situated differently.
	• **Individualistic model**: The malfunction of human reasoning is mainly due to deficits in an individual's reasoning capabilities.	• **Systemic model**: The malfunction of human reasoning is mainly due to the contaminated epistemic environment or lack of support for reasoning to function in its optimal situation [64, 57].

Weina Jin, Ghassan Hamarneh. The Responsible XAI Framework:
Operationalizing Justice Assumptions in the Technical Design and Evaluation Practices of Explainable AI (XAI)

Critical Technical Practice

Demonstration in the explainable AI task

Table 1: A checklist summary of the Responsible XAI framework proposed in Section 3.

1. **Basic properties of XAI**
☐ **P1: Power check.** XAI is to check and shift power from the profit-oriented sector to users and society, especially marginalized and mostly impacted communities.
☐ **P2: User-centred XAI.** XAI should support users' intended goals of using AI explanation and users' reasoning process.
☐ **P3: Bounded functionality.** The XAI/AI system has scopes, weaknesses, and limitations that can or cannot be overcome by technical advances.
☐ **P4: Rejectable AI.** The associated AI system of the XAI technique and its data collection are rejectable.
☐ **P5: Unnecessary tradeoff.** Explainability does not inevitably trade off with performance.

Weina Jin, Ghassan Hamarneh. The Responsible XAI Framework:
Operationalizing Justice Assumptions in the Technical Design and Evaluation Practices of Explainable AI (XAI)

Critical Technical Practice Power-sensitive interventions

1.	**Making power visible, explicable, and checked**
☐	Embed power analysis in AI technical practices, i.e., asking "for whom?" and "who can be put at risk?" in major and minor design choices in AI
☐	Embed power analysis in the routine of ethics review and ethics analysis of AI techniques
☐	Unite just power from the AI community and civil society to take collective actions to check and balance the current dominant power in AI
2.	**Reframing AI and AI ethics for justice**
☐	Reframe AI developmental goals to include the mitigation of harms, in addition to technical innovation (analogous to building a good car, not a good engine only)
☐	Reframe AI ethics as the necessary adversarial guardrail for the healthy development of AI (analogous to regarding the adversarial role of the brake not as a tradeoff with the engine)
3.	**Encoding ethics as technical and scientific methodologies in AI practice**
☐	Critically reflect and examine the taken-for-granted practices and their assumptions in AI, and institutionalize critical examination as a major AI research agenda
☐	Investigate in equivalent amount of resources and attention to algorithmic limitation analysis as the attention on algorithmic performance analysis, and institutionalize limitation analysis as the AI technical evaluation standard
☐	Explicitly encode ethical values and perspectives as technical standards for an AI task or subfield

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

Critical Technical Practice: A methodology

Co-develop the method of baking ethics and responsibility into the DNA of a technique

- Module 1: Ethical and Responsible Technical Practices: Theories and Goals
  - Challenges of operationalizing ethics principles and theories into technical practices
  - **Diagnose** the root causes based on ethics theories
- Module 2: **Methods** of critical technical practice and analysis, covering the entire technical development pipeline, including:
  - Technical development agendas, problem identification, motivation, and formulation
  - Technical design assumptions
  - Technical evaluation
  - Limitation analysis
  - Deployment

Critical Technical Practice: A methodology

critically inspect assumptions and justifications in common technical practices

Goals of the CTP working group

1. Conduct critical examinations and improve the critical reflections of group members' own technical work or technical field
    - Outcomes: some inspiration, alternative perspectives of understanding technology, side projects, or improvements in your existing projects, etc.
2. Interested group members work together to summarize our critical technical practice experience/processes as formal methods that can be taught and delivered to technical and non-technical audiences
    - **Research**: summarize practice and theorize the CTP methodology
    - **Practice**: Develop tools, tutorials, and courses for CS students, technical audiences, users/stakeholders, and the public

Tentative format of the CTP working group

- Group members bring each doubts and skepticism in each own research and technical practice into critical discussion and feedback
- Each time we present some examples of critical analysis, regarding the technical problem formation, design, evaluation, deployment.

Duration

- Summer 2026, can be extended depending on participants' interest
- Meet weekly or biweekly