Towards End-User-Centered Explainable Artificial Intelligence

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

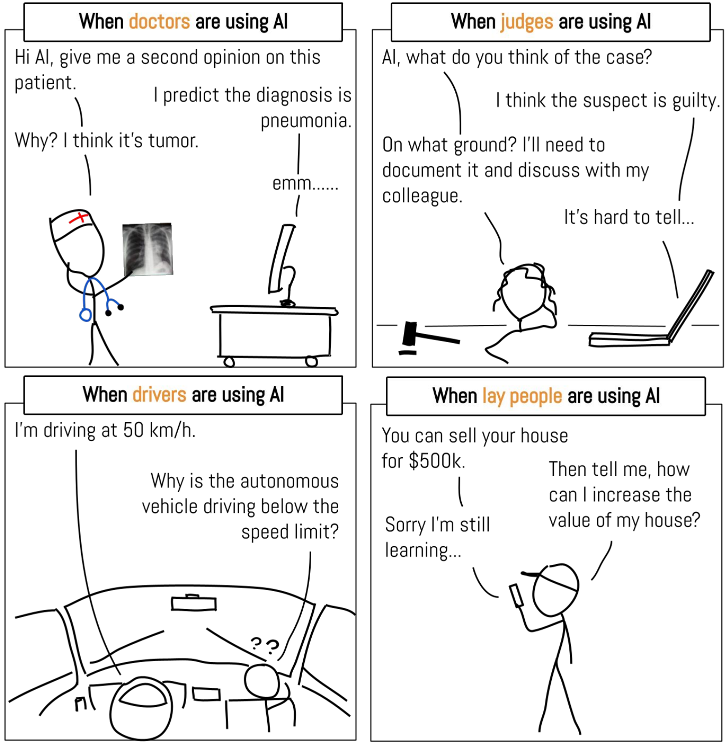
Team:  **Weina Jin**

Advisors:  **Ghassan Hamarneh**  **Xiaoxiao Li**

HAMARNEH
Medical Image Analysis Research Group

SFU SCHOOL OF COMPUTING SCIENCE

The promise of interpretable/explainable AI (XAI)



Explainable AI — Explaining AI decisions in human-understandable ways [1]

Can explainability grant AI models

accountability

trustworthiness

utility to users

?

[1] F. Doshi-Velez, B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017

Prior evaluations on the effectiveness of XAI in end-users' tasks

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

"How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

Osbert Bastani
University of Pennsylvania
obastani@seas.upenn.edu

~~trustworthiness~~

Explanations cannot help users detect potential model biases [2]

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo MIT CSAIL
Michael Muelly Stanford
Hal Abelson MIT CSAIL
Been Kim Google Research

~~accountability~~

Explanations worsen physicians' task performance [3]

ARTICLE Open Access

How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection

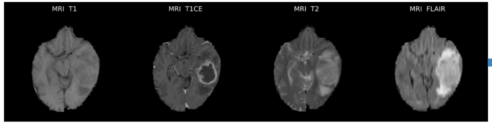
Maia Jacobs[1], Melanie F. Pradier[1], Thomas H. McCoy Jr.[2,3], Roy H. Perlis[2,3], Finale Doshi-Velez[1] and Krzysztof Z. Gajos[1]

~~utility to users~~

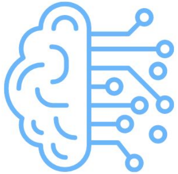
How can we make the AI explanations work as they are supposed to **?**

[1] Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. AIES, 2020
[2] Adebayo, J., Muelly, M., Abelson, H., and Kim, B. Post hoc explanations may be ineffective for detecting unknown spurious correlation. ICLR, 2022
[3] Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 2021.

Co-design XAI with clinical end-users



Input MRI

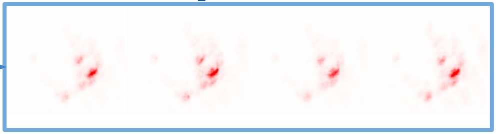


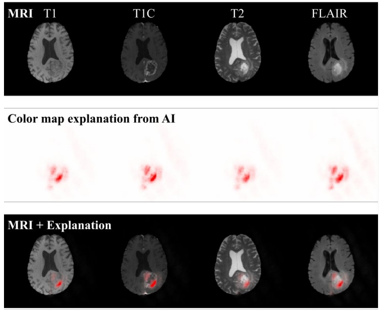
AI prototype

AI's suggestion:

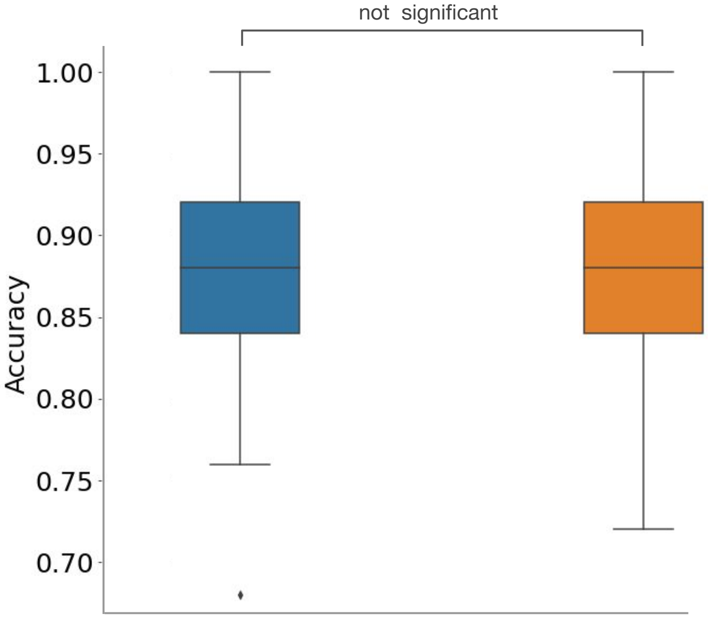
Grade 4 glioblastoma

AI's explanation:



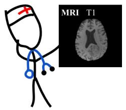


W. Jin, M. Fatehi, R. Guo, G. Hamarneh. Evaluating the Clinical Utility of Artificial Intelligence Assistance and its Explanation on the Glioma Grading Task. 2022. https://doi.org/10.1101/2022.12.07.22282726

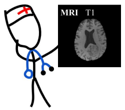


Doctor assisted by AI prediction

Doctor assisted by AI prediction + explanation



Grade 4 tumor



Grade 4 tumor

Evaluating the utility of XAI to clinical end-users

Quantitative results

AI explanations are
not helpful
to improve doctors' task performance

~~utility to users~~

But why ?

W. Jin, M. Fatehi, R. Guo, G. Hamarneh. Evaluating the Clinical Utility of Artificial Intelligence Assistance and its Explanation on the Glioma Grading Task. 2022. https://doi.org/10.1101/2022.12.07.22282726

Evaluating the utility of XAI to clinical end-users

Qualitative results

“

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.

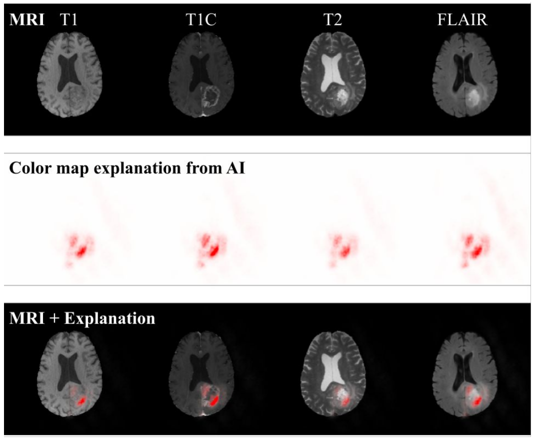
…These color maps were totally useless **without text, without any context or explanation**, like those details. The color maps were just pretty, but they didn’t explain anything.

– Neurosurgeon #3

“

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



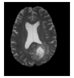
W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023

1. XAI ignores users' explanation 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:

Medical image



→ **Human-interpretable feature** → **Human-interpretable reasoning** based on the features → Clinical decision

Tumor grade 4

Human-interpretable feature:

- Feature location
- Feature description

E.g: contrast-enhanced region of the tumor

Explanation is relevant to clinical decision-making

Guideline 2

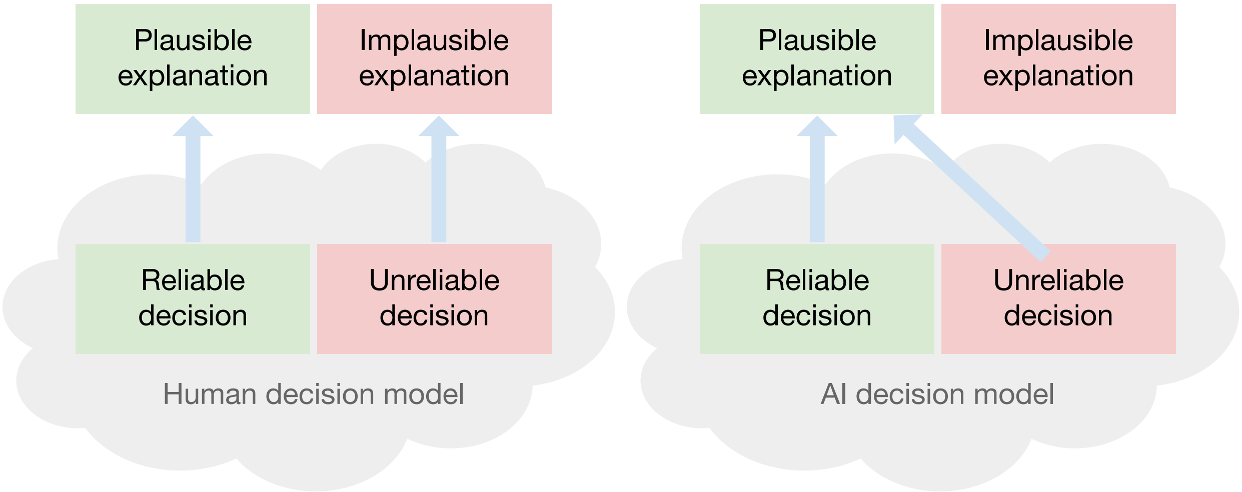
Clinical relevance

What is the human process to incorporate explanations into decision process ?

2. XAI ignores users' communication norm with explanations

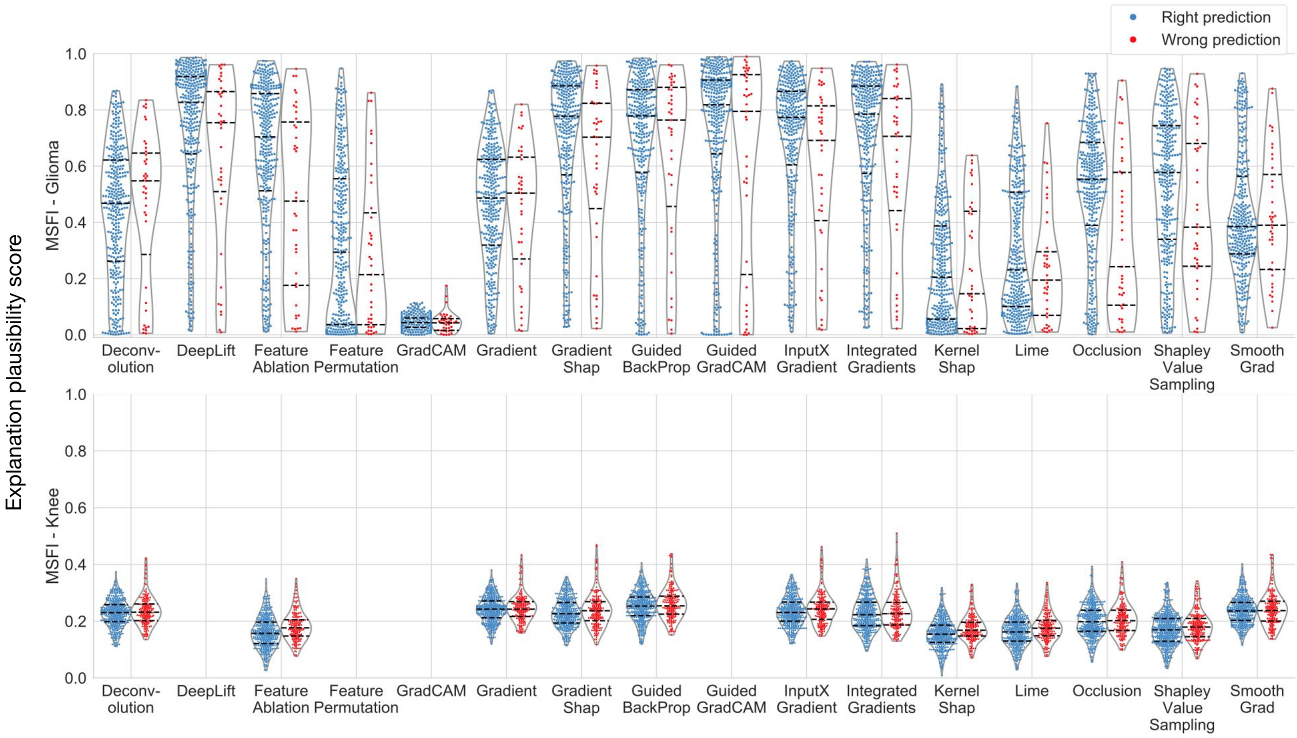
> "That (explanation) is kind of **an internal validation** to me. I want to see scenarios where it (AI) doesn't work, and I can tell that. So this is reassuring to me that, like they **(AI) can make a mistake, and I can call it out, I can determine the mistake**.
>
> – Neurosurgeon #1

Human explanation **AI explanation**



[1] **W. Jin**, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023
[2] **W Jin**, X Li, G Hamarneh. Rethinking AI explainability and plausibility. 2023.

2. XAI ignores users’ communication norm with explanations



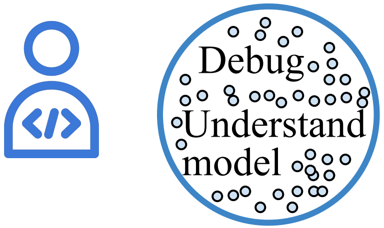
Human judgment on explanation plausibility can reveal decision quality

Guideline 4

Informative plausibility

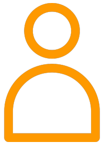
W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023

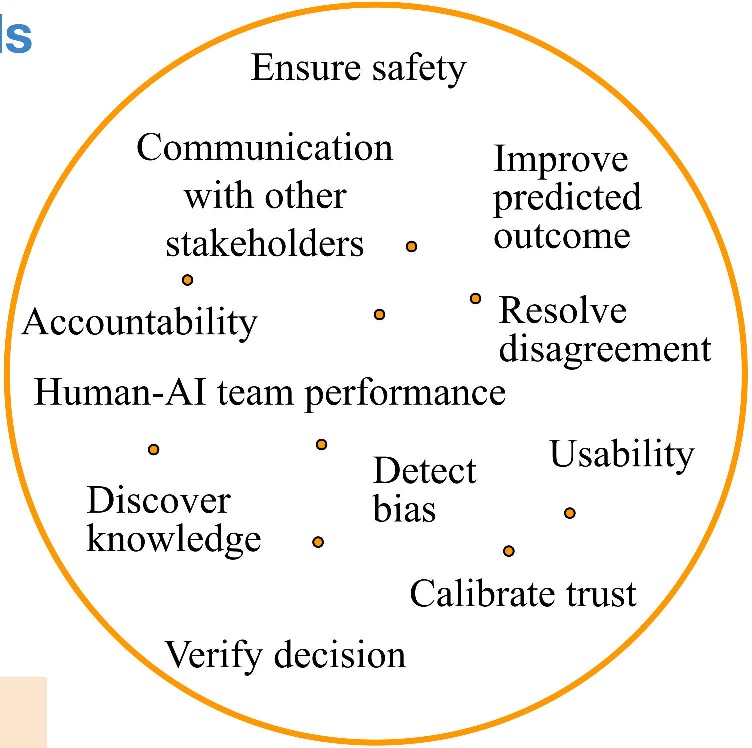
3. XAI ignores users' explanation goals



Technical users' XAI problem space

XAI algorithm





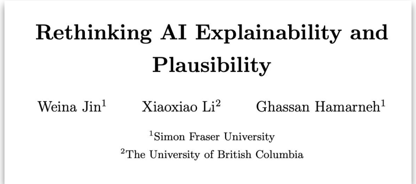
Technical-centered **vs** Human-centered

End-users' XAI problem space

XAI algorithm

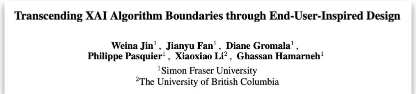
[1] **W Jin**, J Fan, D Gromala, P Pasquier, G Hamarneh. Invisible users: Uncovering end-users' requirements for explainable ai via explanation forms and goals, 2023. Arxiv: 302.06609
[2] **W Jin**, J Fan, D Gromala, P Pasquier, X Li, G Hamarneh. Transcending XAI algorithm boundaries through end-user-inspired design. arxiv: 2208.08739

How to combat the biases against end-users' values?


Rethinking AI Explainability and Plausibility

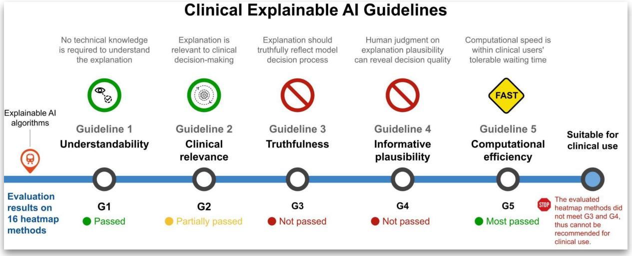
Weina Jin[1] Xiaoxiao Li[2] Ghassan Hamarneh[1]

[1]Simon Fraser University
[2]The University of British Columbia


Transcending XAI Algorithm Boundaries through End-User-Inspired Design

Weina Jin[1], Jianyu Fan[1], Diane Gromala[1], Philippe Pasquier[1], Xiaoxiao Li[2], Ghassan Hamarneh[1]

[1]Simon Fraser University
[2]The University of British Columbia



Raise awareness

Set proper end-user-centered evaluation criteria

Practical tools to support technical specification with end-users

[1] **W Jin**, X Li, G Hamarneh. Rethinking AI explainability and plausibility. 2023.
[2] **W Jin**, J Fan, D Gromala, P Pasquier, X Li, G Hamarneh. Transcending XAI algorithm boundaries through end-user-inspired design. arxiv: 2208.08739
[3] **W. Jin**, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Medical Image Analysis, 2023
[4] **W Jin**, J Fan, D Gromala, P Pasquier, G Hamarneh. Invisible users: Uncovering end-users' requirements for explainable AI via explanation forms and goals, 2023. Arxiv: 302.06609
[5] **W Jin**, J Fan, D Gromala, P Pasquier, G Hamarneh. EUCA: the End-User-Centered Explain- able AI Framework. arXiv:2102.02437, 2021

Unconscious biases in technology and how we combat it

Technology is not value-neutral,
because decisions that shape technology embed values [1].

Unconscious biases are mainly due to:

1. Significant differences in the availability of facts and information

   Diversified perspectives

2. Taking conventions/common practice for granted without critical inspection

   Unconventional thinking

The value and importance of diversity

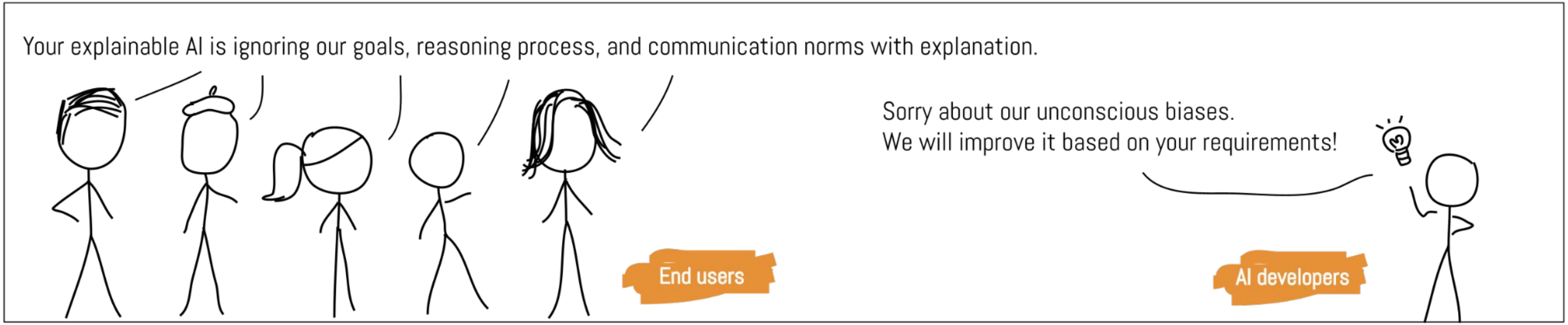
> "Algorithms are designed by people, and **people embed their unconscious biases in algorithms**. It's rarely intentional—but this doesn't mean we should let data scientists off the hook. It means we should be critical about and vigilant for the things we know can go wrong. If we assume discrimination is the default, then we can design systems that work toward notions of equality. [2]

[1] R. J. Whelchel, "Is Technology Neutral?," IEEE Technology and Society Magazine, 1986, doi: 10.1109/MTAS.1986.5010049.
[2] Meredith Broussard, "Popular Doesn't Mean Good," in Artificial Unintelligence: How Computers Misunderstand the World, MIT Press, 2018, pp.149-160.

Thank you!

Towards End-User-Centered Explainable Artificial Intelligence

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



Summary of how XAI ignores end-users' values

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

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

[1] **W Jin**, X Li, G Hamarneh. Rethinking AI explainability and plausibility. 2023.
[2] **W Jin**, X Li, M Fatehi, G Hamarneh. Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis, 2023
[3] **W Jin**, J Fan, D Gromala, P Pasquier, X Li, G Hamarneh. Transcending XAI algorithm boundaries through end-user-inspired design. arxiv: 2208.08739

Prior evaluations on the effectiveness of XAI in end-users' tasks

Technical-centered
vs.
Human-centered

User's trust in AI can easily be manipulated by AI explanations [1]

"How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

Osbert Bastani
University of Pennsylvania
obastani@seas.upenn.edu

~~trustworthiness~~

Explanations cannot help users detect potential model biases [2]

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo
MIT CSAIL

Michael Muelly
Stanford

Hal Abelson
MIT CSAIL

Been Kim
Google Research

~~accountability~~

Explanations worsen physicians' task performance [3]

ARTICLE Open Access

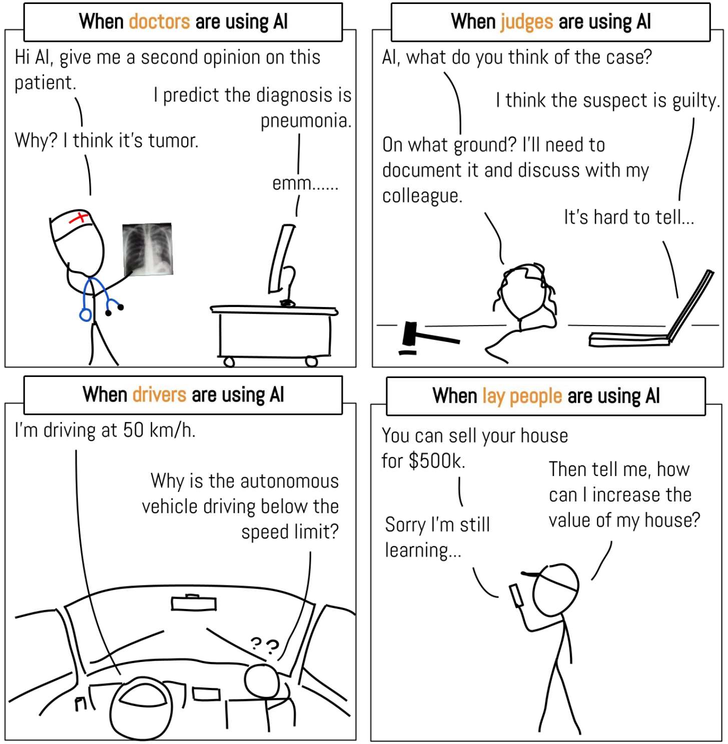
How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection

Maia Jacobs[1], Melanie F. Pradier[1], Thomas H. McCoy Jr.[2,3], Roy H. Perlis[2,3], Finale Doshi-Velez[1] and Krzysztof Z. Gajos[1]

~~utility to users~~

[1] Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. AIES, 2020
[2] Adebayo, J., Muelly, M., Abelson, H., and Kim, B. Post hoc explanations may be ineffective for detecting unknown spurious correlation. ICLR, 2022
[3] Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 2021.

The promise of interpretable/explainable AI (XAI)



Explainable AI

Explaining AI decisions in human-understandable ways [1]

?

The "promise" of XAI

- Trustworthiness
- Accountability
- Improving task performance

[1] F. Doshi-Velez, B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017

Prior evaluations on the effectiveness of XAI in end-users' tasks

User's trust in AI can easily be manipulated by AI explanations [1]

"How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

Osbert Bastani
University of Pennsylvania
obastani@seas.upenn.edu

Explanations cannot help users detect potential model biases [2]

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo MIT CSAIL
Michael Muelly Stanford
Hal Abelson MIT CSAIL
Been Kim Google Research

Explanations worsen physicians' task performance [3]

ARTICLE Open Access

How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection

Maia Jacobs[1], Melanie F. Pradier[1], Thomas H. McCoy Jr.[2,3], Roy H. Perlis[2,3], Finale Doshi-Velez[1] and Krzysztof Z. Gajos[1]

Explainable AI

The "promise" of XAI

- ~~Trustworthiness~~
- ~~Accountability~~
- ~~Improving task performance~~

[1] Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. AIES, 2020
[2] Adebayo, J., Muelly, M., Abelson, H., and Kim, B. Post hoc explanations may be ineffective for detecting unknown spurious correlation. ICLR, 2022
[3] Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 2021.

Towards End-User-Centered Explainable Artificial Intelligence

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

Team:  **Weina Jin**

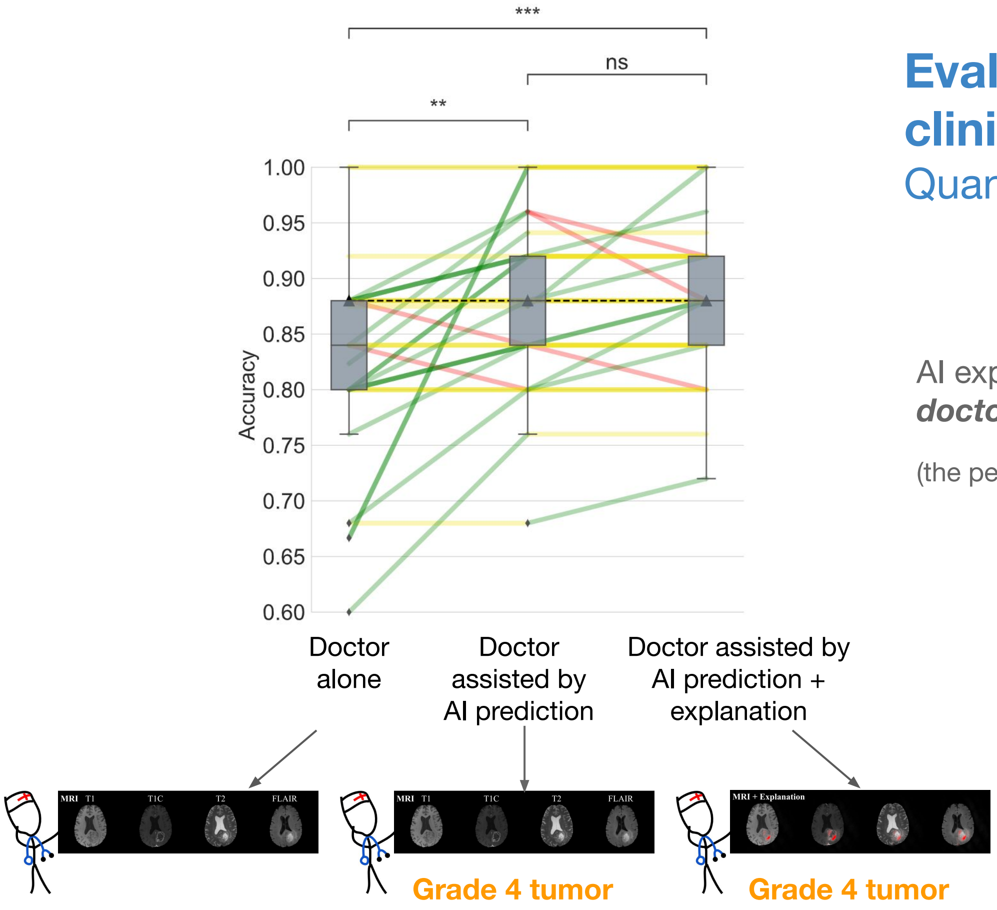
Advisors:  **Ghassan Hamarneh**  **Xiaoxiao Li**

HAMARNEH
Medical Image Analysis
Research Group

SFU SCHOOL OF COMPUTING SCIENCE

Evaluating the utility of XAI to clinical end-users

Quantitative results



AI explanations are **not helpful to improve *doctor-AI team performance***

(the performance that surpass doctor or AI alone)

W. Jin, M. Fatehi, R. Guo, G. Hamarneh. Evaluating the Clinical Utility of Artificial Intelligence Assistance and its Explanation on the Glioma Grading Task. 2022. https://doi.org/10.1101/2022.12.07.22282726

Towards End-User-Centered Explainable Artificial Intelligence

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

Team:  **Weina Jin**

Advisors:  **Ghassan Hamarneh**

 **Xiaoxiao Li**

HAMARNEH
Medical Image Analysis Research Group

SFU SIMON FRASER UNIVERSITY

The effectiveness of feature attribution methods and its correlation with automatic evaluation scores

il.com

Daeyoung Kim
KAIST
kimd@kaist.ac.kr

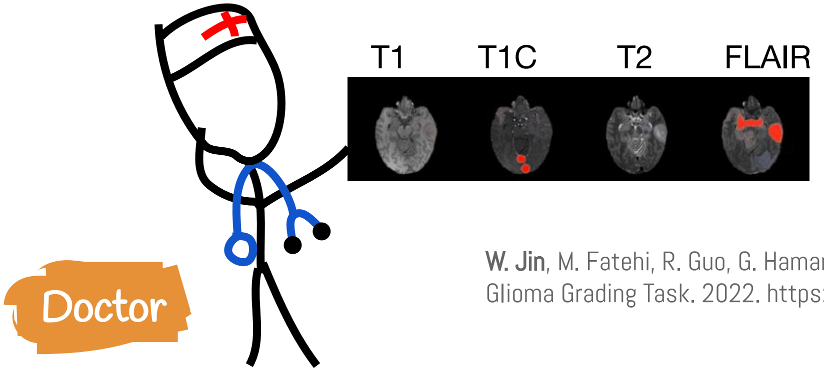
Anh Nguyen[*]
Auburn University
nh.ng8@gmail.com

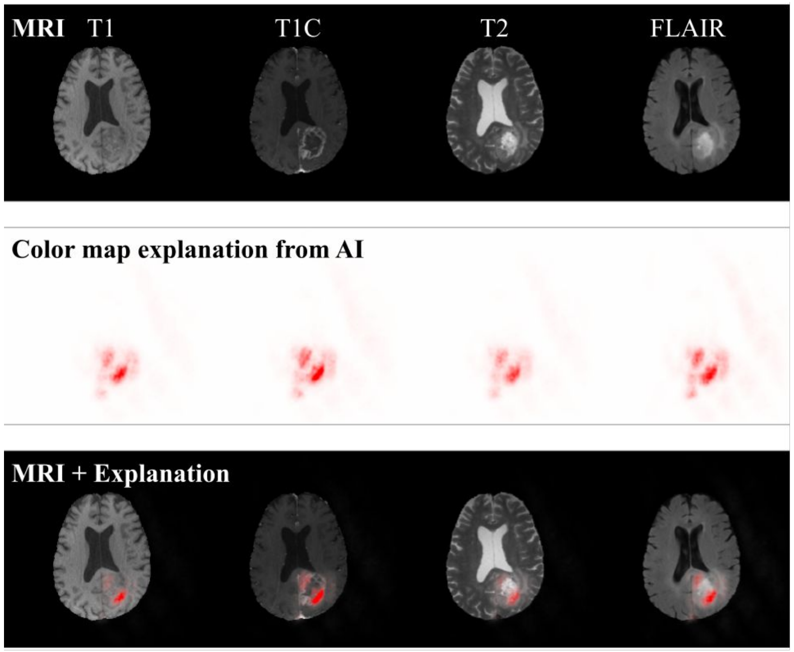
"

Feature attribution is **surprisingly not more effective** than showing humans nearest training-set examples. On a harder task of fine-grained dog categorization, presenting **attribution maps** to humans does not help, but instead **hurts the performance of human-AI teams** compared to AI alone.

Nguyen, G., Kim, D., Nguyen, A. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. NeurIPS. 2021.

Evaluating the utility of XAI to clinical end-users





W. Jin, M. Fatehi, R. Guo, G. Hamarneh. Evaluating the Clinical Utility of Artificial Intelligence Assistance and its Explanation on the Glioma Grading Task. 2022. https://doi.org/10.1101/2022.12.07.22282726

Jacobs et al. *Translational Psychiatry* (2021)11:108
https://doi.org/10.1038/s41398-021-01224-x

Translational Psychiatry

ARTICLE Open Access

How machine-learning recommend
influence clinician treatment selecti
example of antidepressant selectio

Maia Jacobs[1], Melanie F. Pradier[1], Thomas H. McCoy Jr.[2,3], Roy H. Perlis[2,3], Final
Krzysztof Z. Gajos[1]

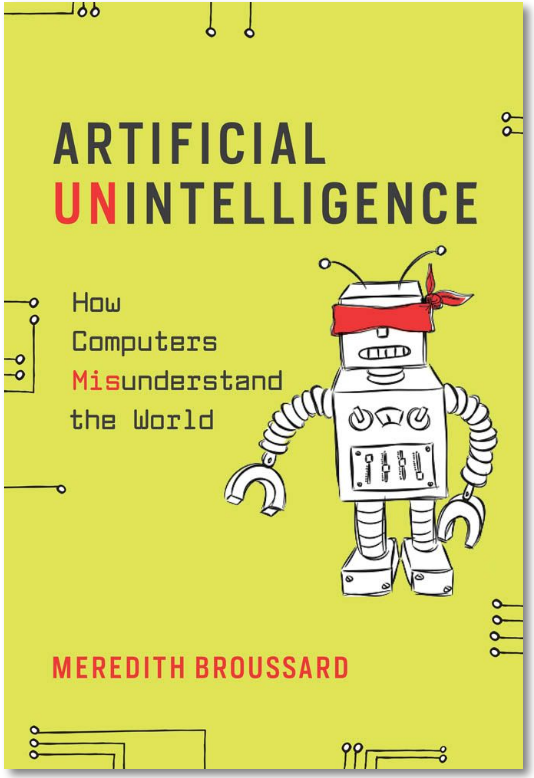
“

Incorrect ML recommendations may adversely impact clinician treatment selections and that **explanations are insufficient for addressing overreliance on imperfect ML algorithms**.

[1] Nguyen, G., Kim, D., Nguyen, A. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. NeurIPS. 2021.
[2] Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational Psychiatry, 2021.

"

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



[1] Meredith Broussard, "Popular Doesn't Mean Good," in Artificial Unintelligence: How Computers Misunderstand the World, MIT Press, 2018, pp.149-160.

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

Most unconscious biases are due to:

1. Significant differences in the availability of facts and information
2. Taking conventions/common practice for granted without critical inspection

The value and importance of diversity

[1] Meredith Broussard, "Popular Doesn't Mean Good," in Artificial Unintelligence: How Computers Misunderstand the World, MIT Press, 2018, pp.149-160.