Stop Using Plausibility as the Criterion for Explainable AI

Weina Jin Medical Image Analysis Lab, School of Computing Science, Simon Fraser University

Xiaoxiao Li Department of Electrical and Computer Engineering, The University of British Columbia
Vector Institute

Ghassan Hamarneh Medical Image Analysis Lab, School of Computing Science, Simon Fraser University



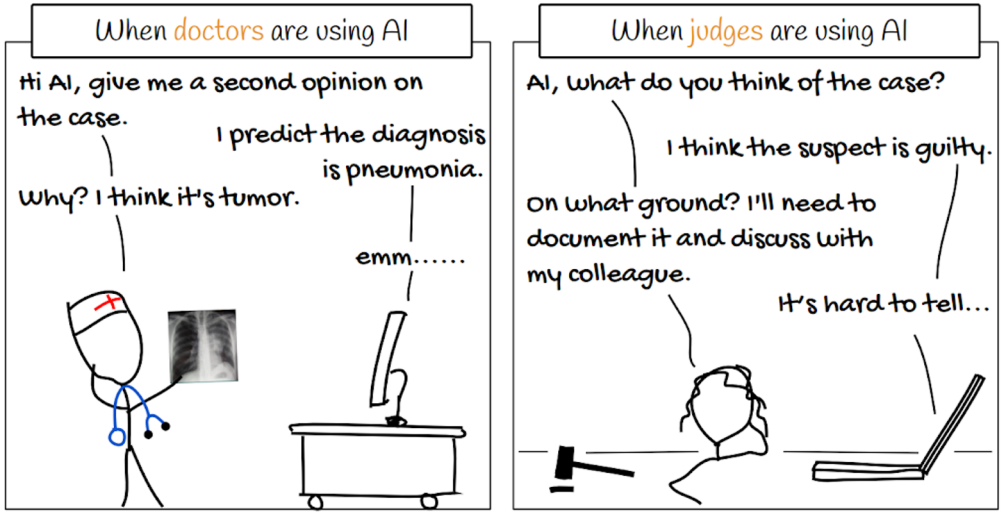




eXplainable AI (XAI)

Explainability, accountability, transparency

AI applications in high-stakes tasks

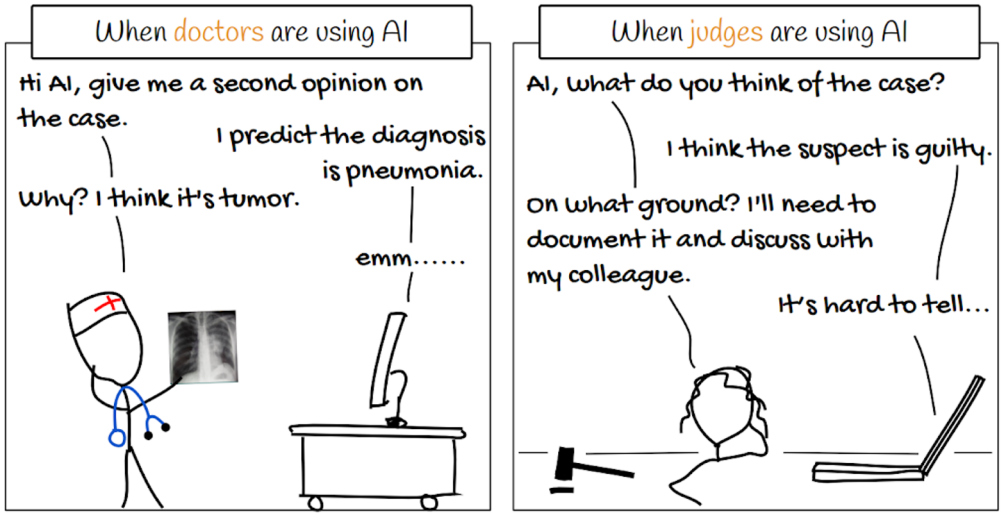


- **Motivation: Responsible evaluation of XAI**

eXplainable AI (XAI)

Explainability, accountability, transparency

AI applications in high-stakes tasks



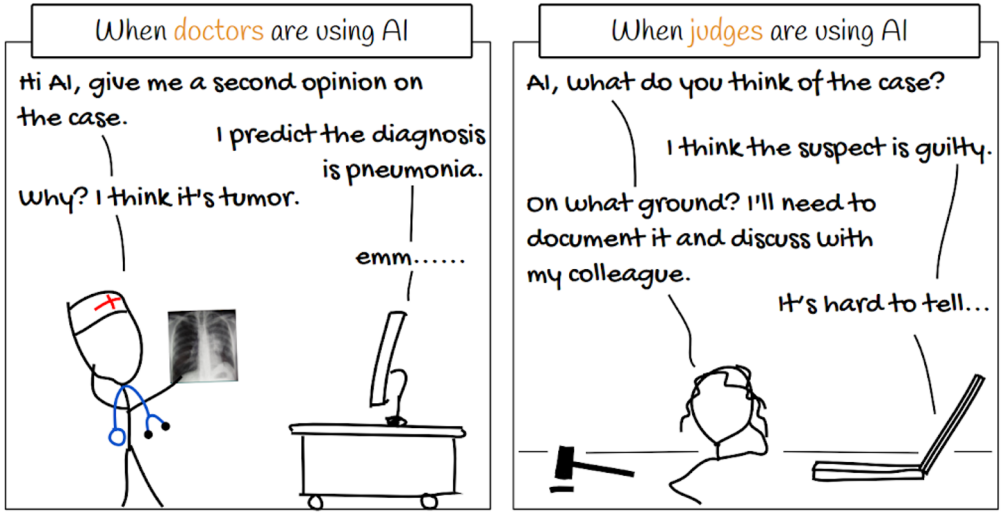
- **Motivation: Responsible evaluation of XAI**

eXplainable AI (XAI)

Rigorous evaluation

Explainability, accountability, transparency

AI applications in high-stakes tasks



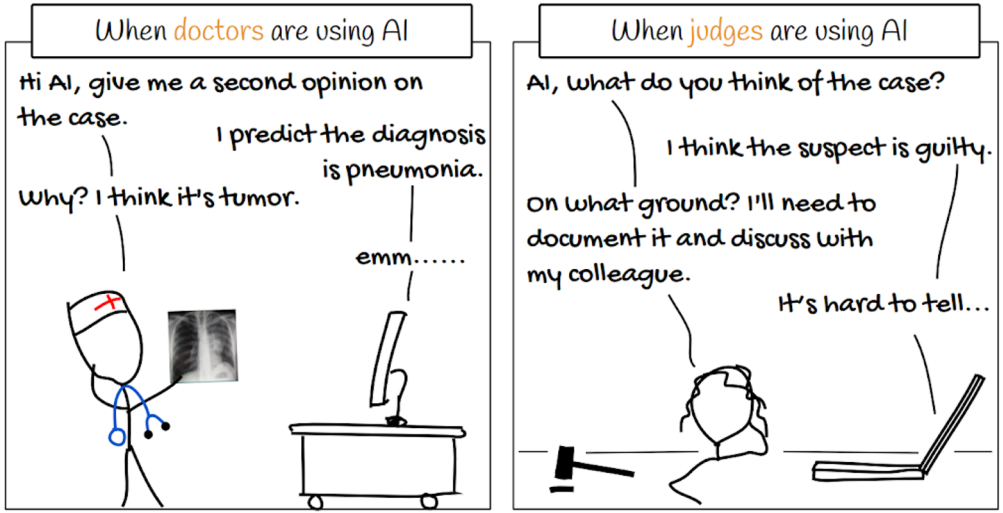
- **Motivation: Responsible evaluation of XAI**

eXplainable AI (XAI)

Rigorous evaluation

Explainability, accountability, transparency

AI applications in high-stakes tasks



- **Motivation: Responsible evaluation of XAI**

eXplainable AI (XAI)

Rigorous evaluation



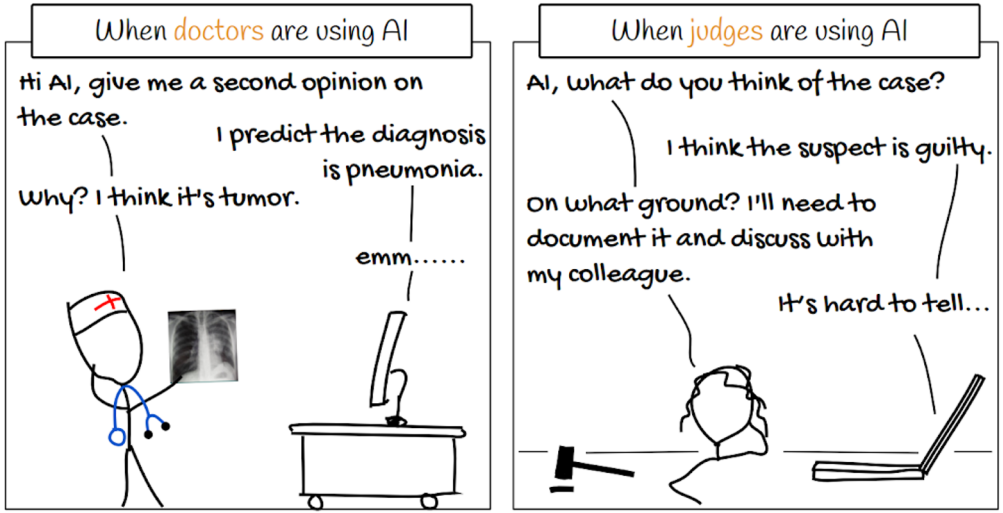
Explainability, accountability, transparency

AI applications in high-stakes tasks



Critical examination of a **main evaluation criterion**

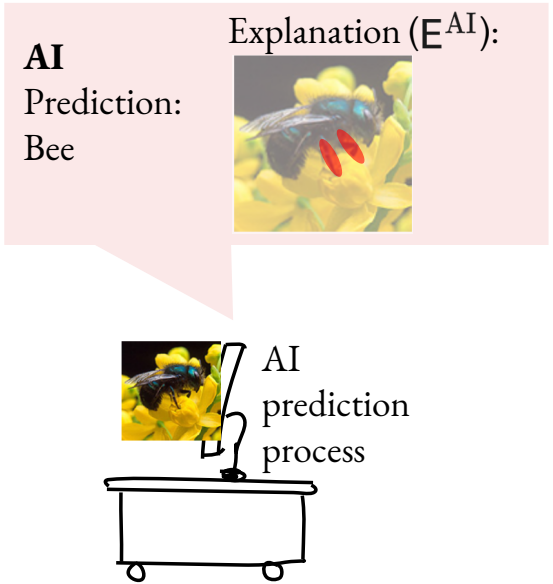
plausibility



● Definition

Plausibility: the similarity between AI and human explanations[1]

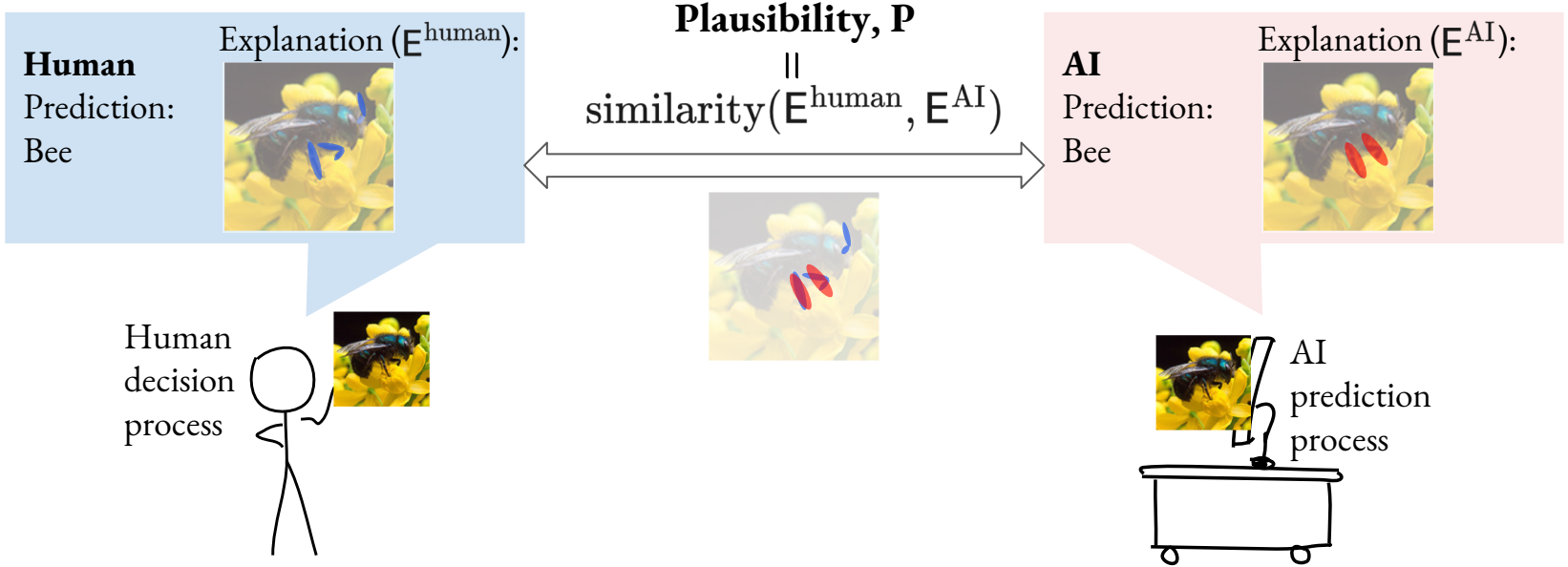
Plausibility, P



[1] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? 2020. ACL

● Definition

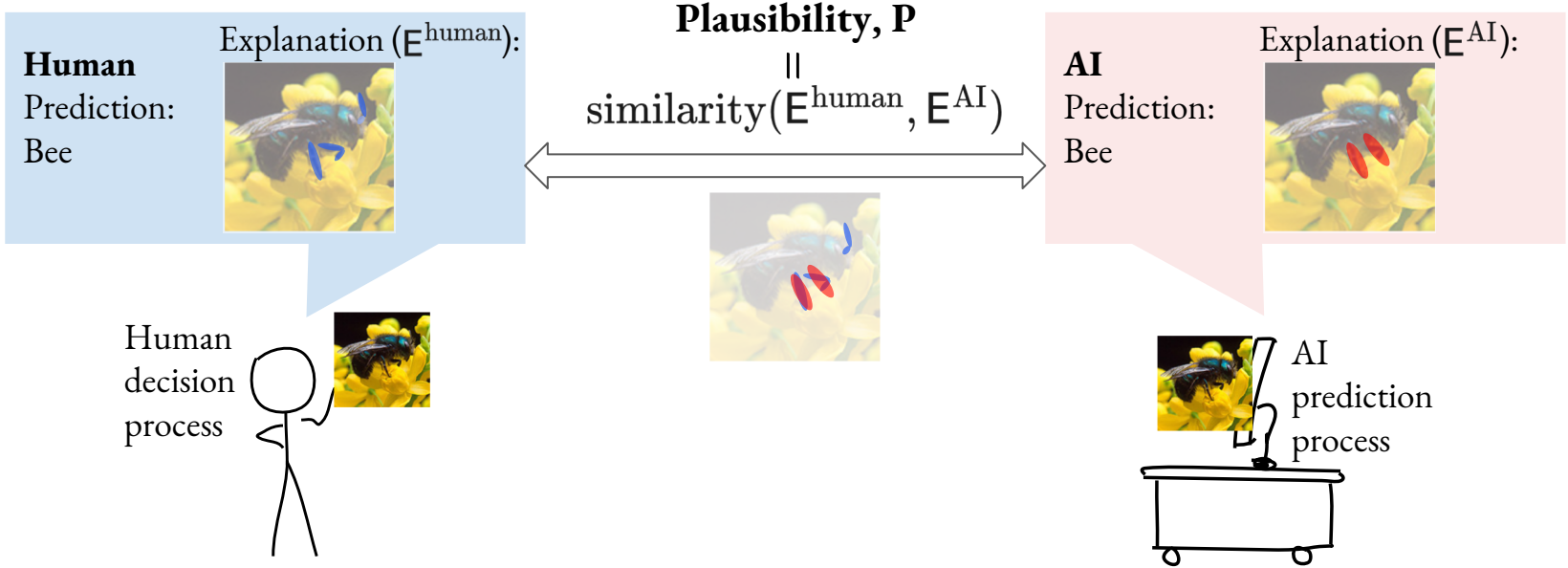
Plausibility: the similarity between AI and human explanations[1]



[1] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? 2020. ACL

The plausibility criterion is widely used in XAI evaluation

- **1/3** of XAI papers published in prominent AI conferences used the plausibility criterion[1].
- The plausibility criterion is common in XAI reviews and programming libraries.



[1] Meike Nauta, et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. 2023. ACM Computing Surveys

Our position

Plausibility, a commonly used evaluation criterion of explainable AI (XAI), is **unscientific** and **unethical**.

Our position

Plausibility, a commonly used evaluation criterion of explainable AI (XAI), is **unscientific** and **unethical**.

The AI and XAI communities should **stop using the plausibility criterion** to evaluate and optimize XAI algorithms.

- Unscientific measure

Plausibility lacks construct validity to measure explainability

AI explanation (the effect)



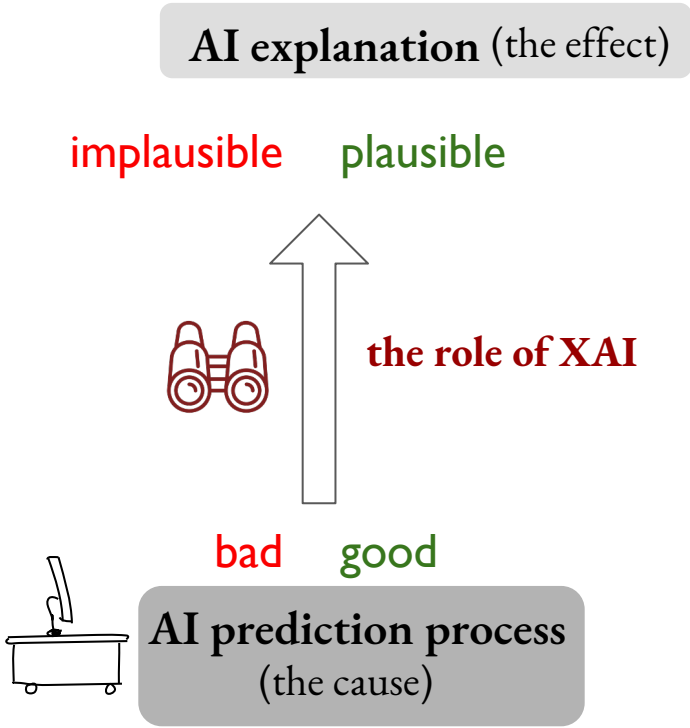
the role of XAI



AI prediction process
(the cause)

- Unscientific measure

Plausibility lacks construct validity to measure explainability



- Unscientific measure

Plausibility lacks construct validity to measure explainability

AI explanation (the effect)

implausible plausible



the role of XAI

bad good



AI prediction process
(the cause)

Audit report (the effect)

negative positive



the role of auditor

bad good



Company's business conduct
(the cause)

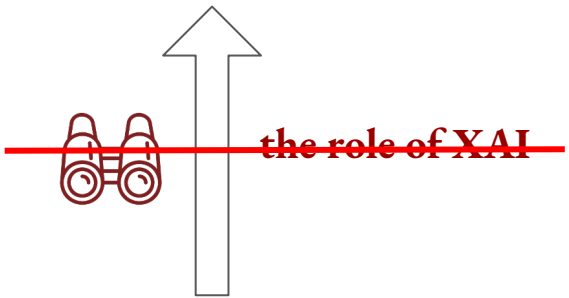
- Unscientific measure

Plausibility lacks construct validity to measure explainability

Incentivize XAI to create more plausible explanations

AI explanation (the effect)

implausible plausible



bad good

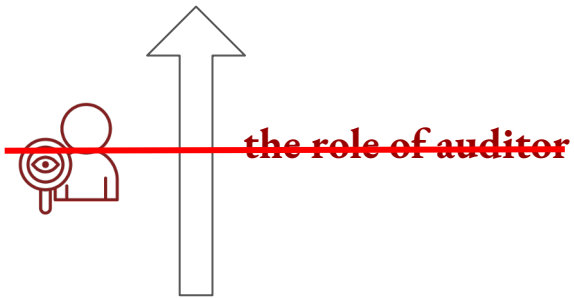


AI prediction process
(the cause)

Incentivize (bribe) auditor to create more positive reports

Audit report (the effect)

negative positive



bad good



Company's business conduct
(the cause)

- Unscientific measure

Plausibility lacks construct validity to measure explainability

AI explanation (the effect)

implausible **plausible**



the role of XAI

bad **good**



AI prediction process
(the cause)

Improving AI model's learning

Audit report (the effect)

negative **positive**



the role of auditor

bad **good**

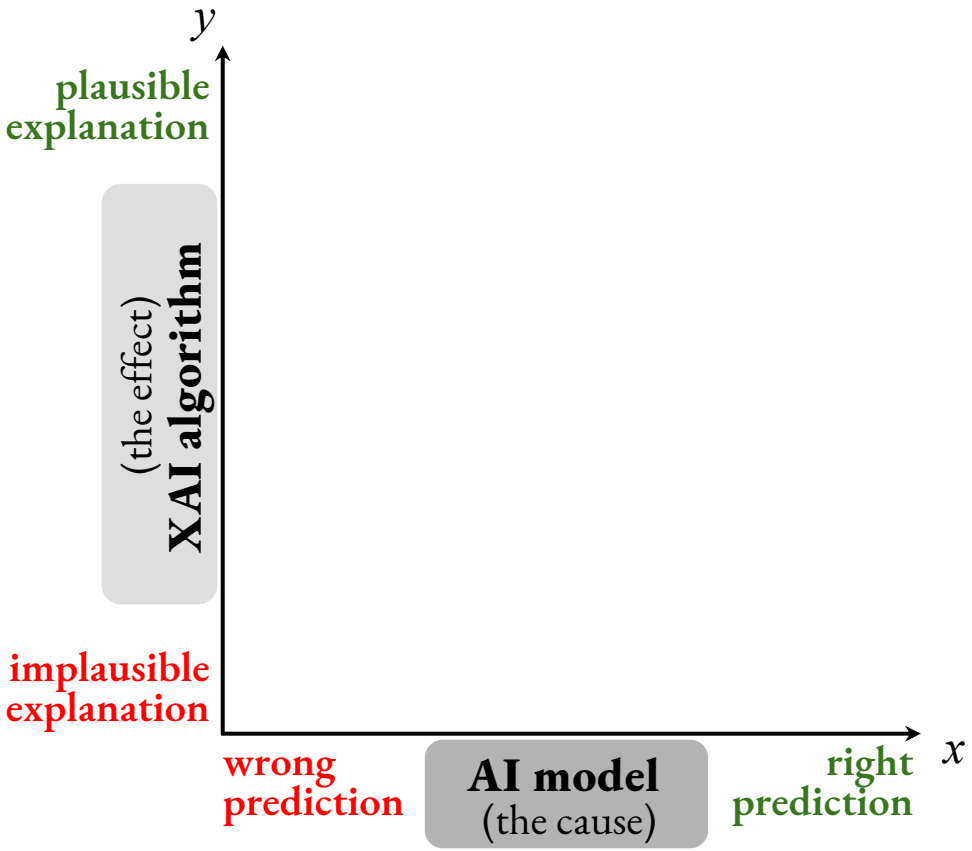


Company's business conduct
(the cause)

Improving the company's conduct

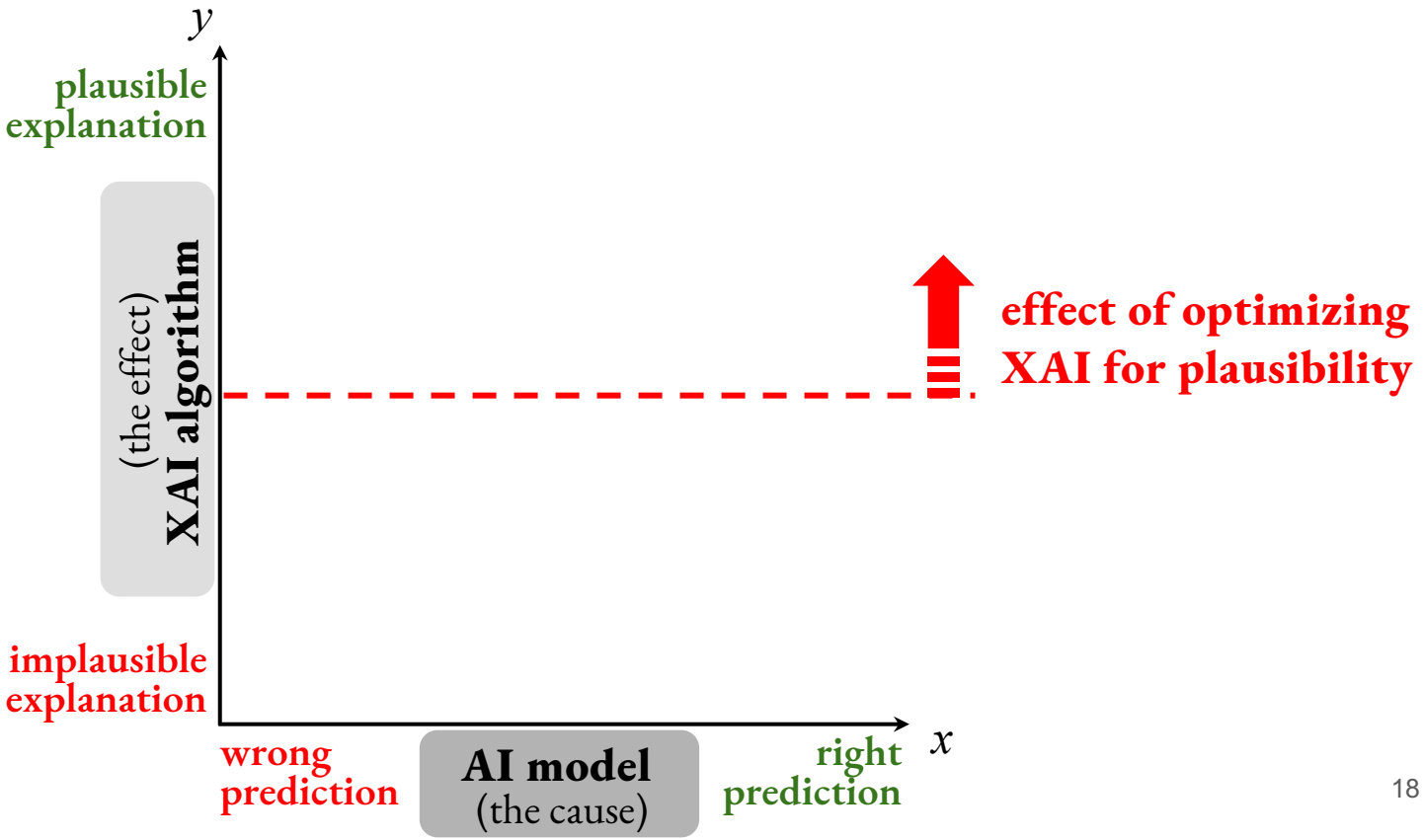
- Unethical consequence

Using the plausibility criterion increases misleading explanations

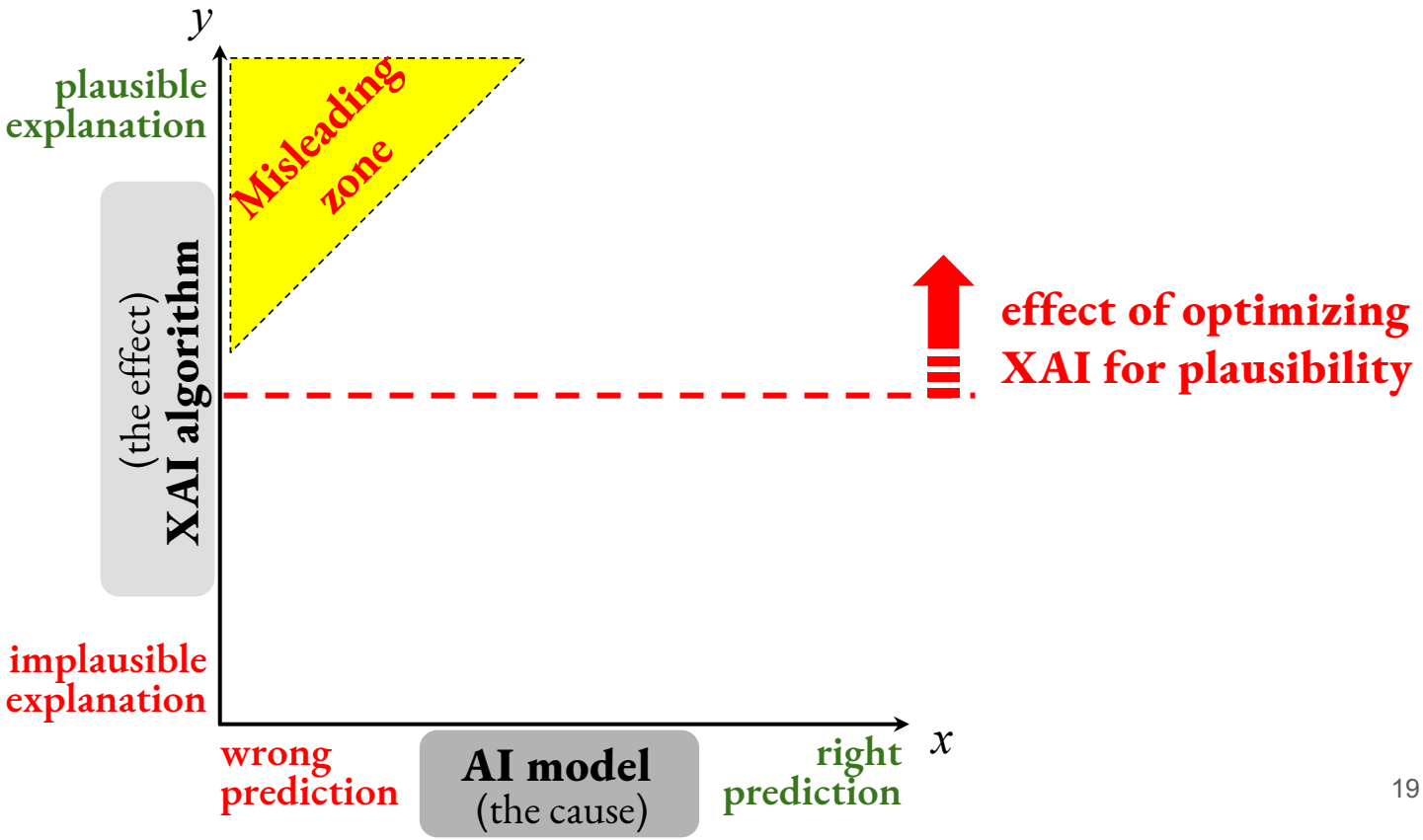


- Unethical consequence

Using the plausibility criterion increases misleading explanations



Using the plausibility criterion increases misleading explanations



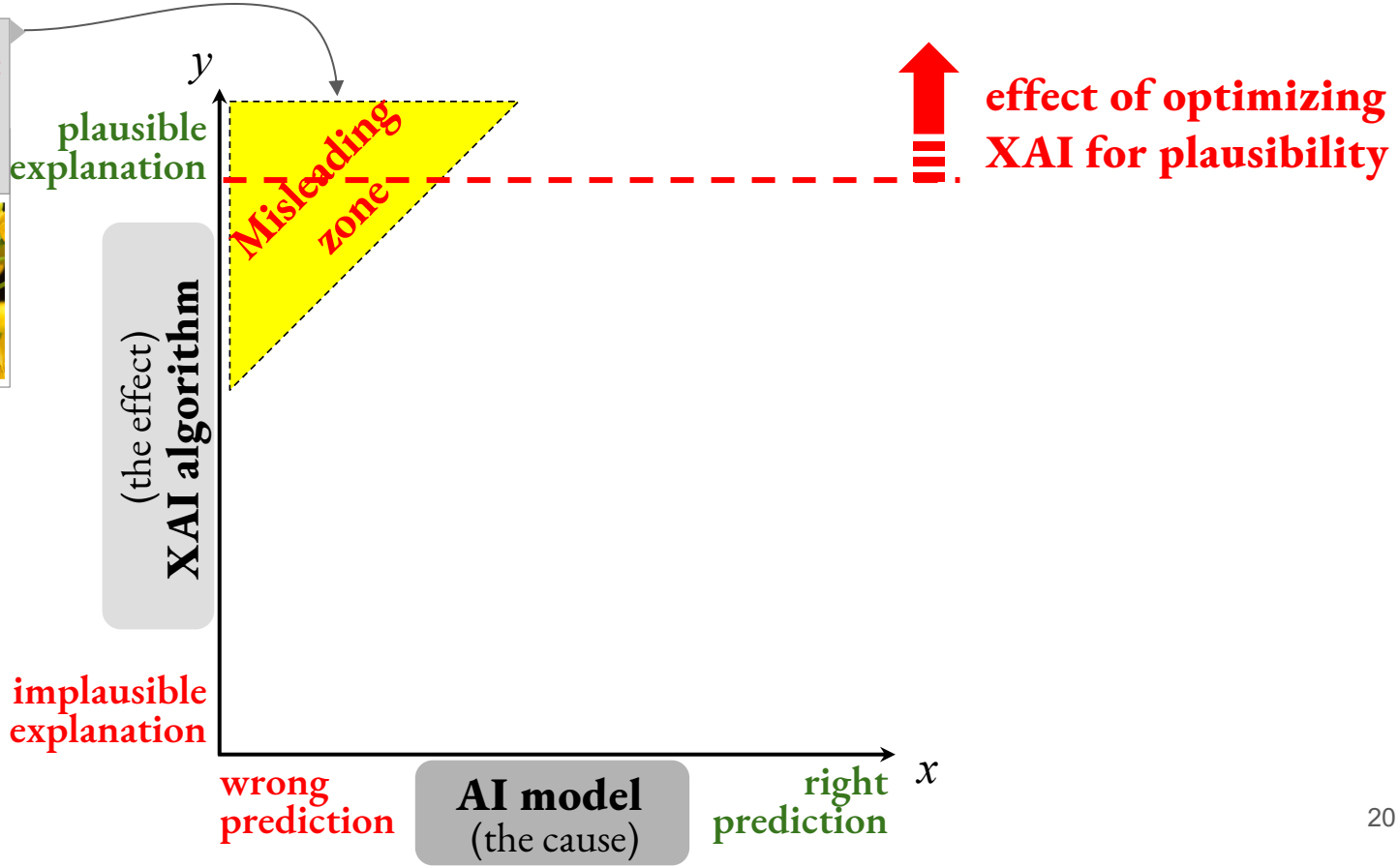
- Unethical consequence

Using the plausibility criterion increases misleading explanations

Misleading explanation:
Plausible reason, wrong prediction

User: *Why do you predict it is a **fly**?*
AI: *Because I identify the discriminative features of big eyes and green body.*





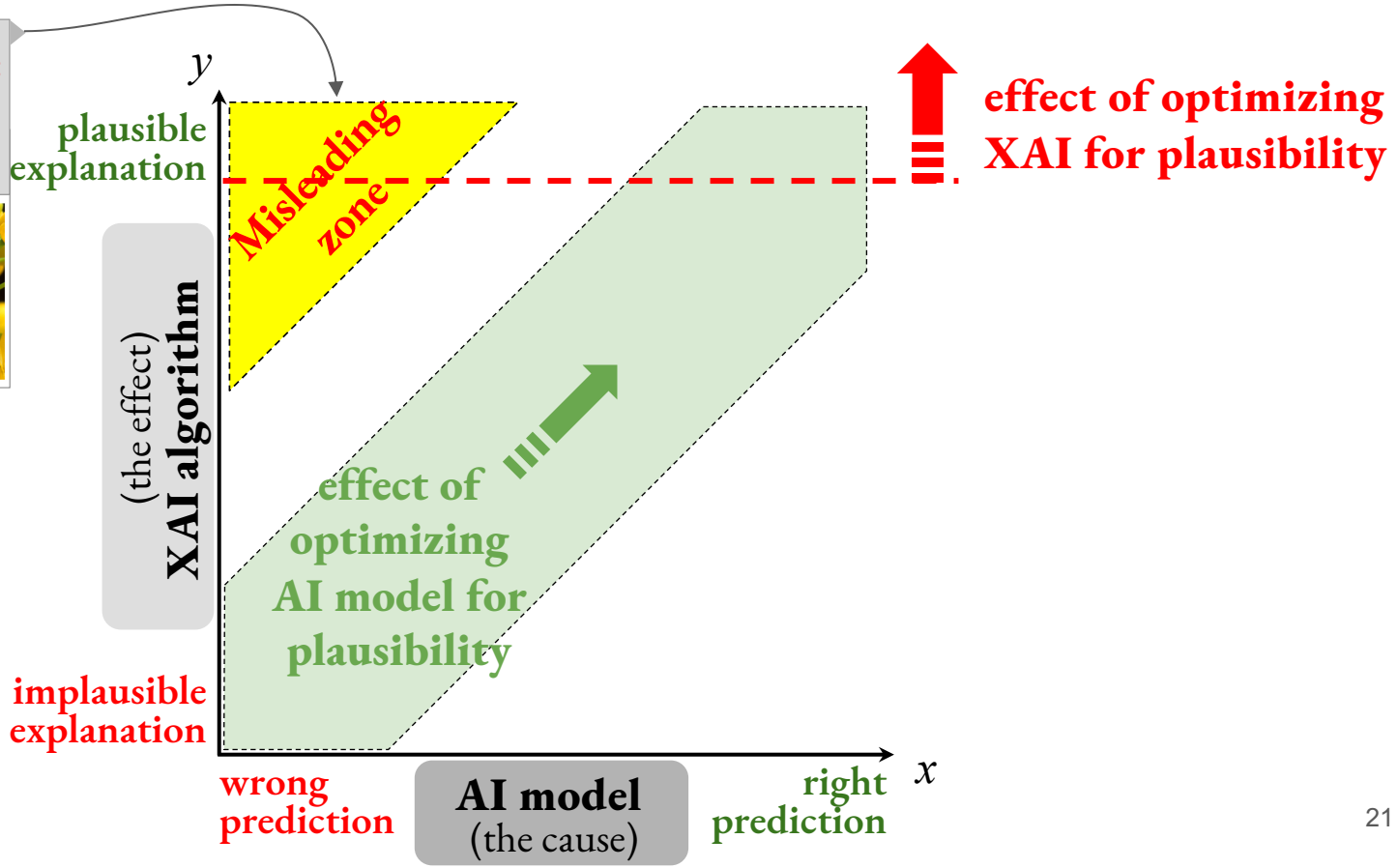
- Unethical consequence

Using the plausibility criterion increases misleading explanations

Misleading explanation:
Plausible reason, wrong prediction

User: *Why do you predict it is a **fly**?*
AI: *Because I identify the discriminative features of big eyes and green body.*





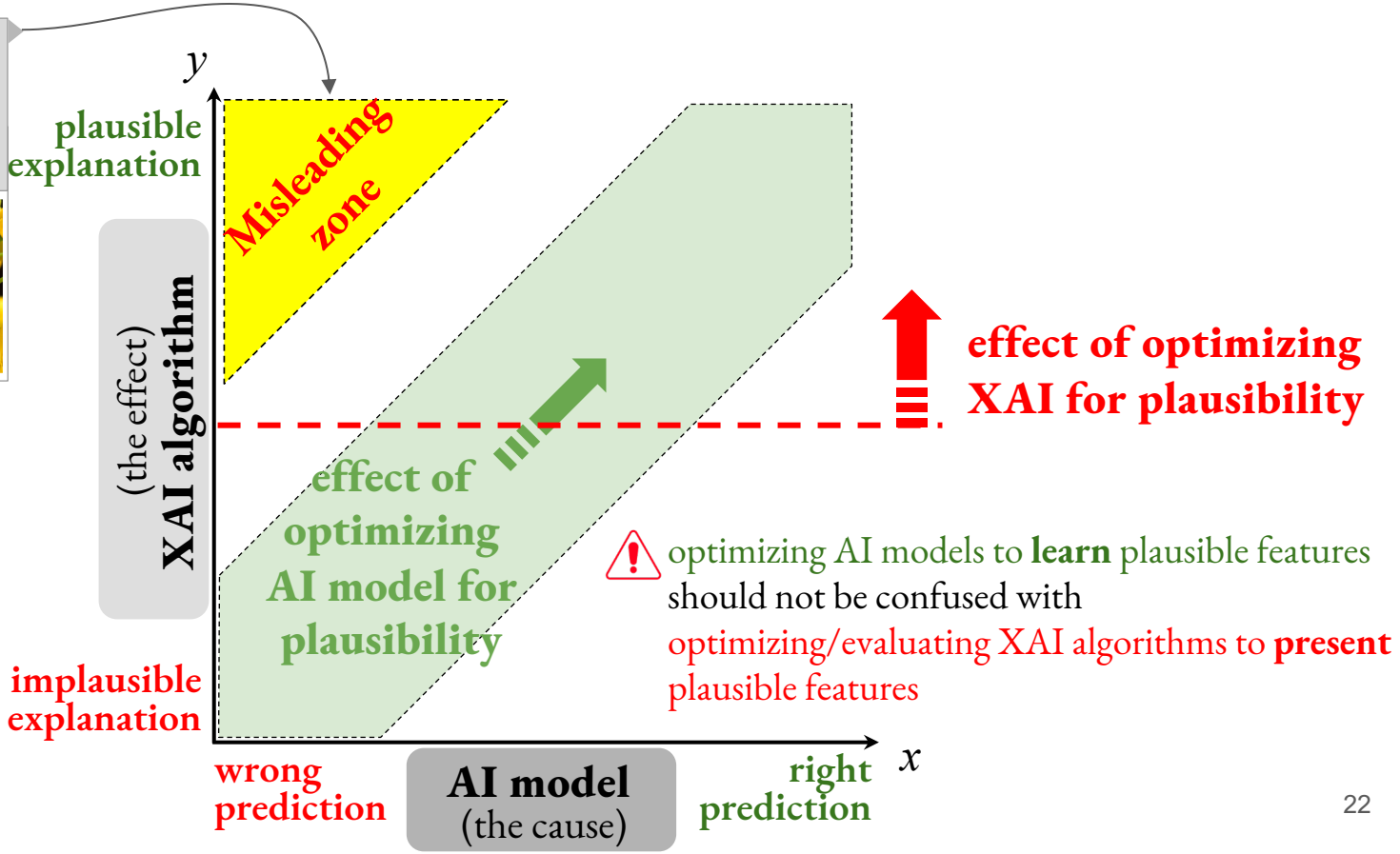
- Unethical consequence

Using the plausibility criterion increases misleading explanations

Misleading explanation:
Plausible reason, wrong prediction

User: *Why do you predict it is a **fly**?*
AI: *Because I identify the discriminative features of big eyes and green body.*





Human explanations are NOT ground truth of AI explanations

Human
Prediction:
Bee

(the effect)
Explanation (E^{human}):

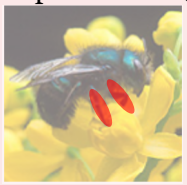


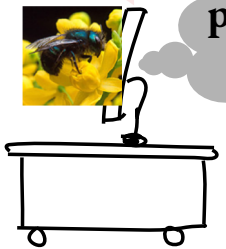
Human decision process
(the cause)



AI
Prediction:
Bee

(the effect)
Explanation (E^{AI}):





AI prediction process
(the cause)

Two independent processes, although their explanation contents may overlap

Consequences of using the plausibility criterion

1. Increase misleading explanations that can manipulate users and destroy users' trust. → We suggest ways to use plausibility properly.

2. Cannot achieve complementary performance: human-AI team outperforms either human or AI. → We identify the theoretical condition of developing (X)AI to achieve complementary performance.

3. Miss other possible approaches of enhancing understandability. → We provide ways to enhance understandability.

● Takeaways

Stop Using Plausibility as the Criterion for Explainable AI

Using the plausibility criterion to evaluate and optimize XAI

⚠ **unscientific** **cannot measure explainability**
"bribing the auditor"

Using human explanations as ground truth for AI explanations

⚠ **unethical** **misleading explanation**

Open question:

What are missing in our community to prevent prevalent mistakes like this? [1]

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