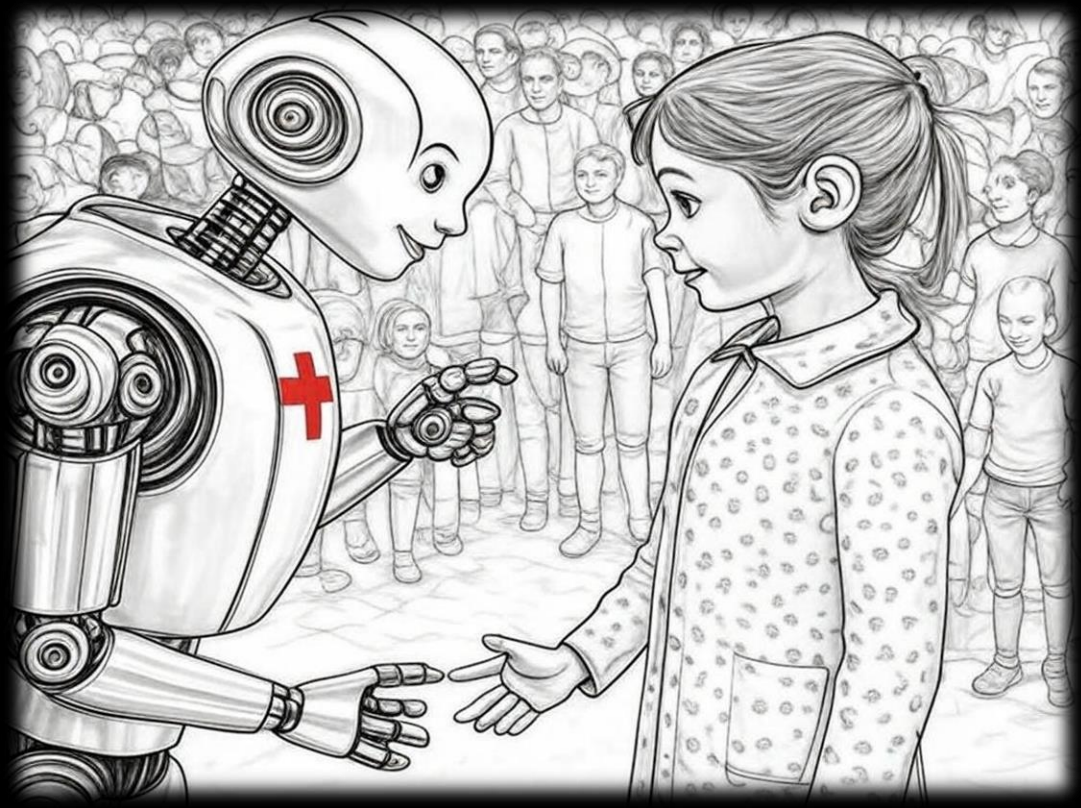
Predictive AI in Behavioral Health and Beyond: Challenges and Lessons Learned

Alka Menon, Ph.D.
Assistant Professor of Sociology
Yale University

Book: AI, Medicine and Society

- With Xenophon Papademetris (Biomedical Informatics and Data Science) and Emily Johns
- Coming out with Cambridge University Press, 2027 or 2028



Results from an interdisciplinary collaboration



Alka Menon
Yale Sociology

Lynn Fiellin
Yale School of Medicine

Xenophon Papademetris
Yale School of Medicine

Christian Kastner
Carnegie Mellon School of Computer Science

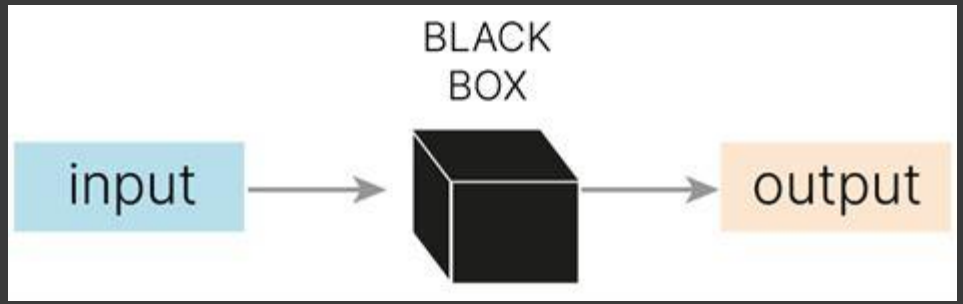
This work is supported by a grant from the National Science Foundation (CISE/CCF 2131466) and by graduate students Nadia Nahar and Zahra Abba Omar.

Opaque AI and accountability

- Arguments for the use of algorithms as more efficient and objective are not new (e.g., Brayne and Christin 2021)
- But the deep learning models behind them are comparatively new
  - Opacity: Don't know what the AI part is doing; no one does
  - Explainability techniques for insight
- Opaque AI systems pose unique problems for accountability
  - Beyond generative AI



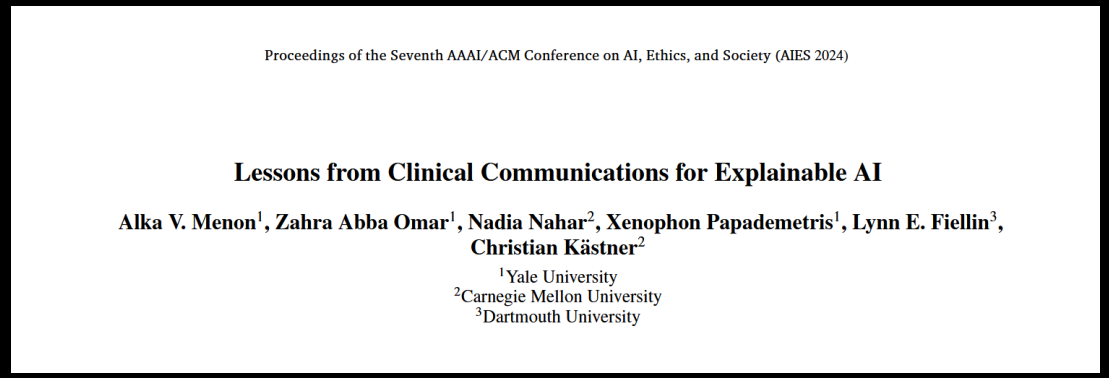
OPENING THE BLACK BOX



[Text]

Transparency and explainability techniques

- Accountability through transparency and system design
- Explainability techniques: Can a variety of stakeholders get into the “black box” of the AI system/ Do they want to?
- Increasing interest in explanations for end users

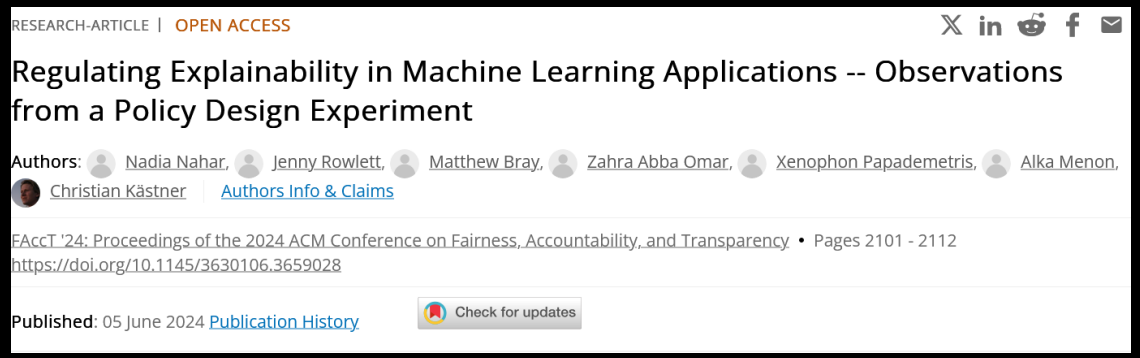


Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Lessons from Clinical Communications for Explainable AI

Alka V. Menon[1], Zahra Abba Omar[1], Nadia Nahar[2], Xenophon Papademetris[1], Lynn E. Fiellin[3], Christian Kästner[2]

[1]Yale University
[2]Carnegie Mellon University
[3]Dartmouth University



RESEARCH-ARTICLE | OPEN ACCESS

Regulating Explainability in Machine Learning Applications -- Observations from a Policy Design Experiment

Authors: Nadia Nahar, Jenny Rowlett, Matthew Bray, Zahra Abba Omar, Xenophon Papademetris, Alka Menon, Christian Kästner | Authors Info & Claims

FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency • Pages 2101 - 2112
https://doi.org/10.1145/3630106.3659028

Published: 05 June 2024 Publication History

Check for updates

Explanations

- **Explanations for AI systems:** Broadly construed, including transparency, disclosure, and inspired by disclosure norms from clinical informed consent
  - For non-technical end users: AI model is in use, and how it made a decision for an individual case
  - What other considerations should there be? Purpose? Audience?
- Explainability (post hoc) techniques: Obtain insight into how an AI makes a prediction and decision, from input to output (e.g., Anchor, SHAP, Lime, etc.)
- Rest of the presentation: Findings from two studies (one failed!) and lessons learned

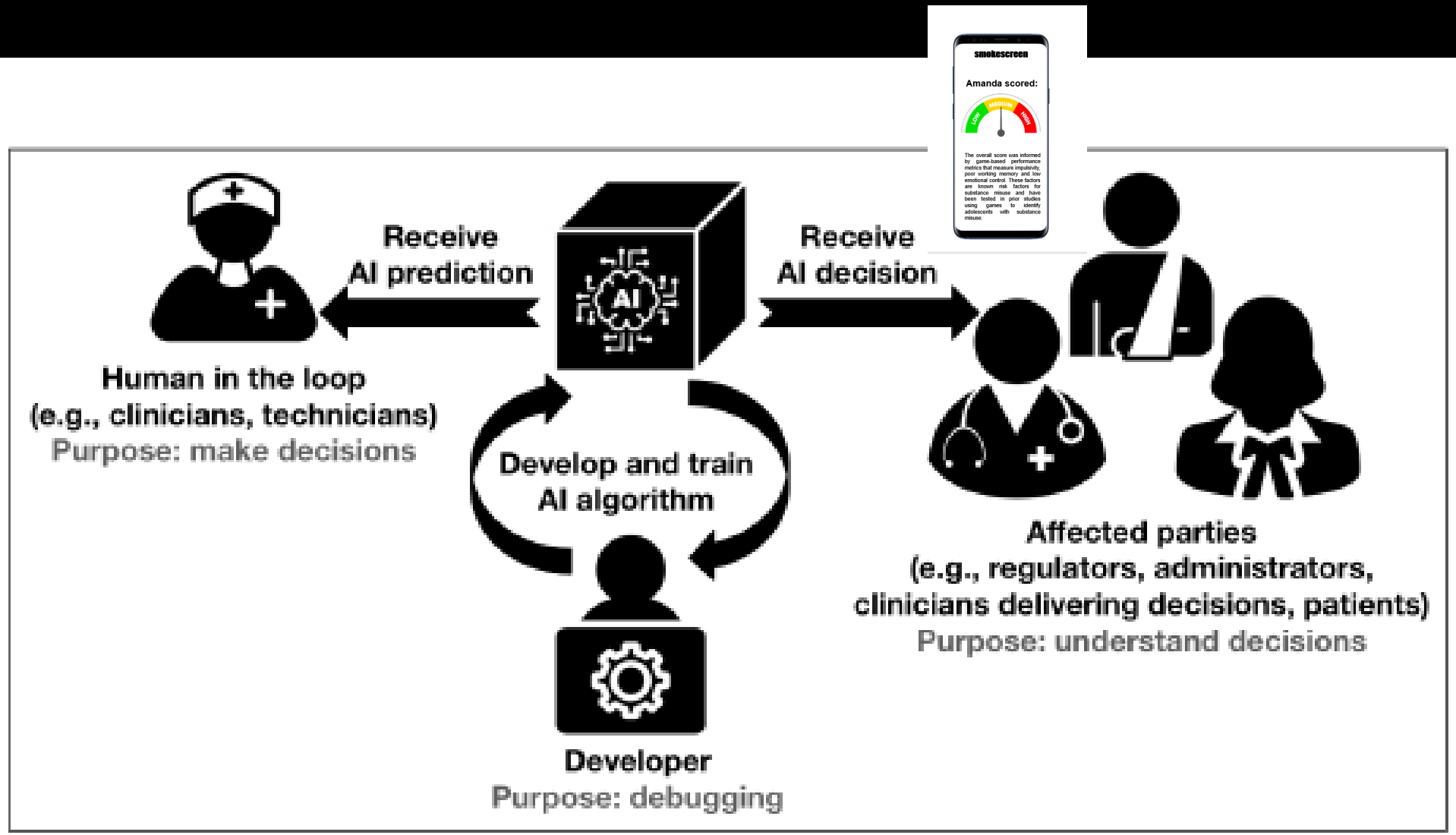


Figure 1: Explanations could be offered at different points in the process and to fulfill different needs

Multiple explanations for different reasons

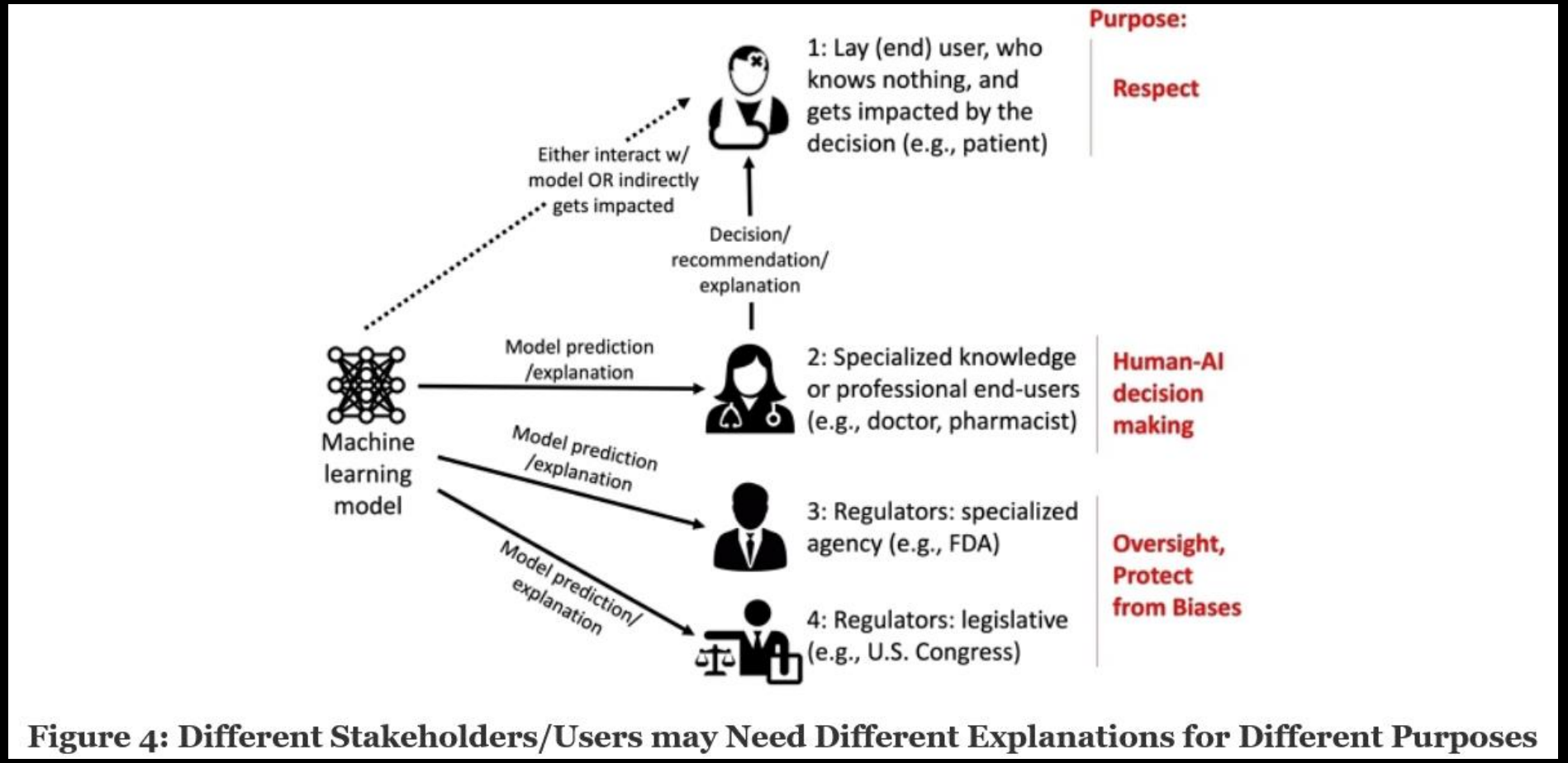


Figure 4: Different Stakeholders/Users may Need Different Explanations for Different Purposes

https://dl.acm.org/doi/fullHtml/10.1145/3630106.3659028

Study 1

Our failure to create a predictive AI screening tool for substance use in teens

Explanations for “low risk” AI systems

SmokeOut, a predictive educational game for teens

“SmokeOut”



- A game designed by public health researchers
- Longitudinal outcome data collected about vaping, smoking, and opioid use
- User play data: cognitive deficits, reaction time, impulsivity
- Plan: AI would generate a substance abuse risk score based on user play

Scenario

Prediction accuracy is found to be on par with existing survey screening tools (e.g., CRAFFT 2.1)

If students received a high risk score, they were referred to school guidance counselor for a session

AI prediction trained based on past game play data combined with adolescent behavioral health outcomes

“Designing Accountable Software System”

- Interviewed 27 stakeholders (parents, schoolteachers, public health researchers, and psychiatrists) for what they wanted in such a system, and specifically about what they wanted for the AI component in 2023
- Computer scientist simultaneously worked with a psychiatrist to sift through the game play data for predictive insights
- Game play data (e.g., reaction time) had significant problems; not enough data points; not generalizable; ultimately not predictive. Designed for one purpose and could not be retooled for a prediction.

Game designed to facilitate multiple scenarios



Study design

- 27 interviews conducted jointly by a computer scientist and sociologist
- Open-ended, beginning with people familiar with the game and expanding to others
- Developed risk score screen mock-ups

Stakeholder type	Number of respondents
Developer/designers (public health experts, coders, software engineering students)	14
End users (teachers/school administrators, education students)	11
Regulators	2
	27

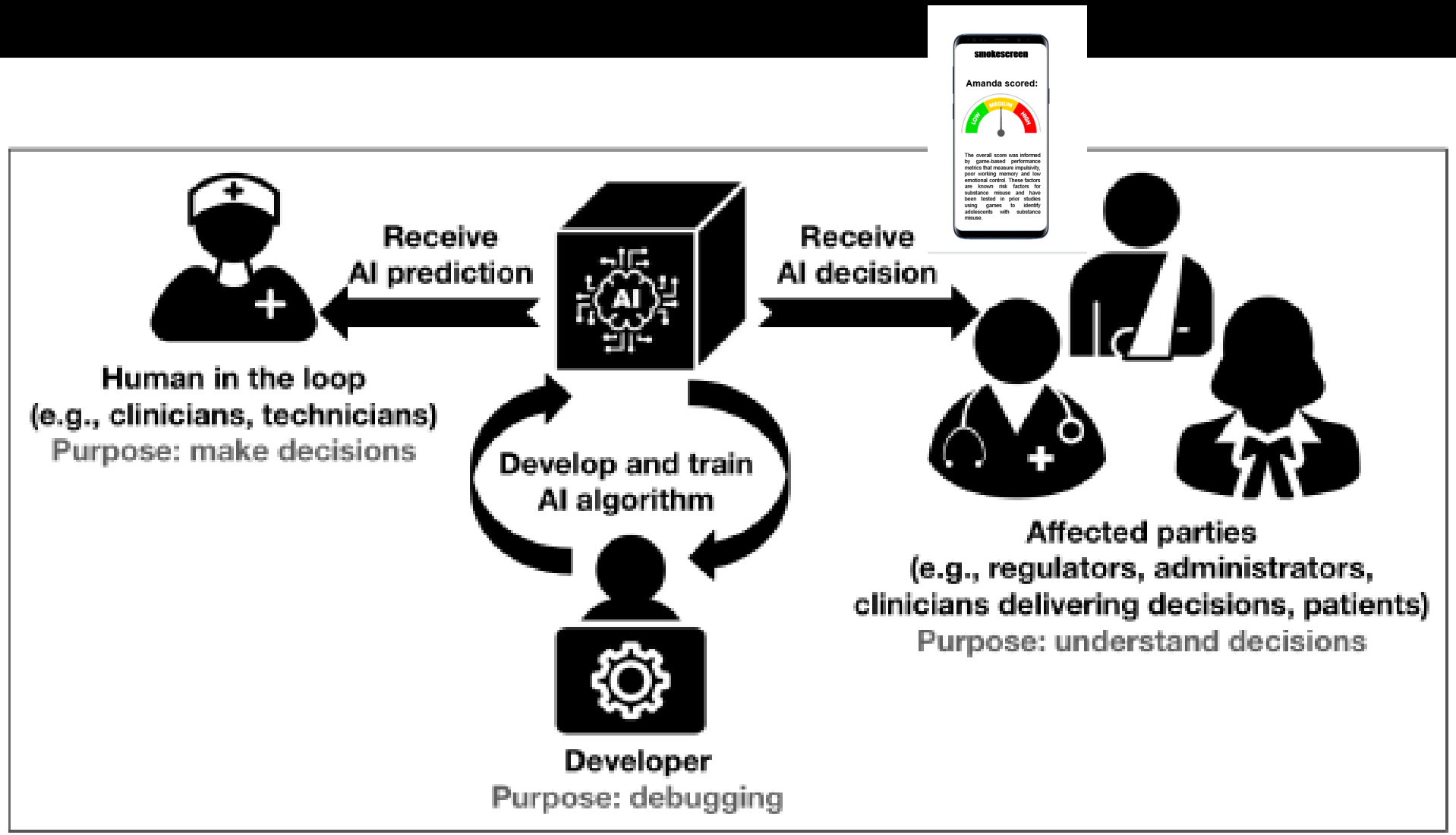
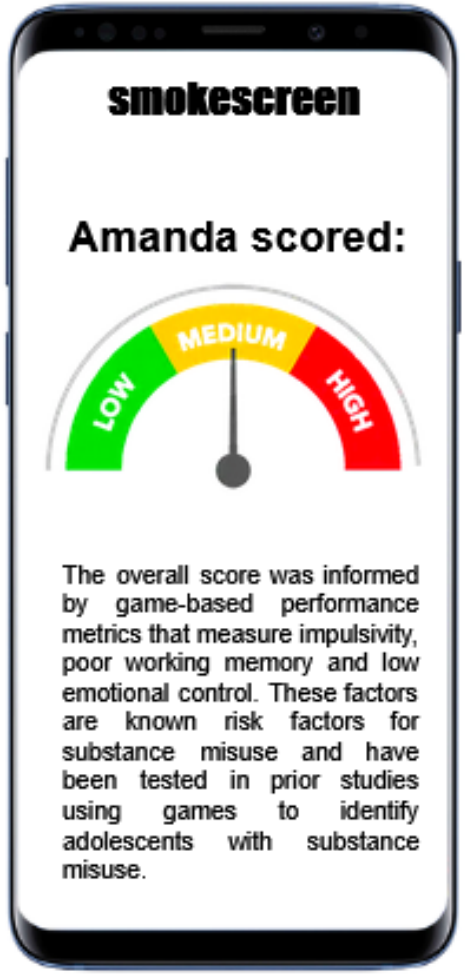


Figure 1: Explanations could be offered at different points in the process and to fulfill different needs


smokescreen
Amanda scored:
LOW
MEDIUM
HIGH
The overall score was informed by game-based performance metrics that measure impulsivity, poor working memory and low emotional control. These factors are known risk factors for substance misuse and have been tested in prior studies using games to identify adolescents with substance misuse.

What information do stakeholders want to know about an AI software system?

Question and answers

- What information do stakeholders want to know about a “low risk” AI software system, and why?
  - Explanations of AI systems function as arguments and justifications for a particular course of action, playing a small role in legitimacy of AI systems.
  - Stakeholders generally want to know that AI is involved. Beyond that, there is little consensus on the information they want to know, including level of (technical) detail, plausibility, conceptual vs. process driven, and who the audience is.

Findings overview

All stakeholders perceive value in conceptual explanations for why a decision was generated.

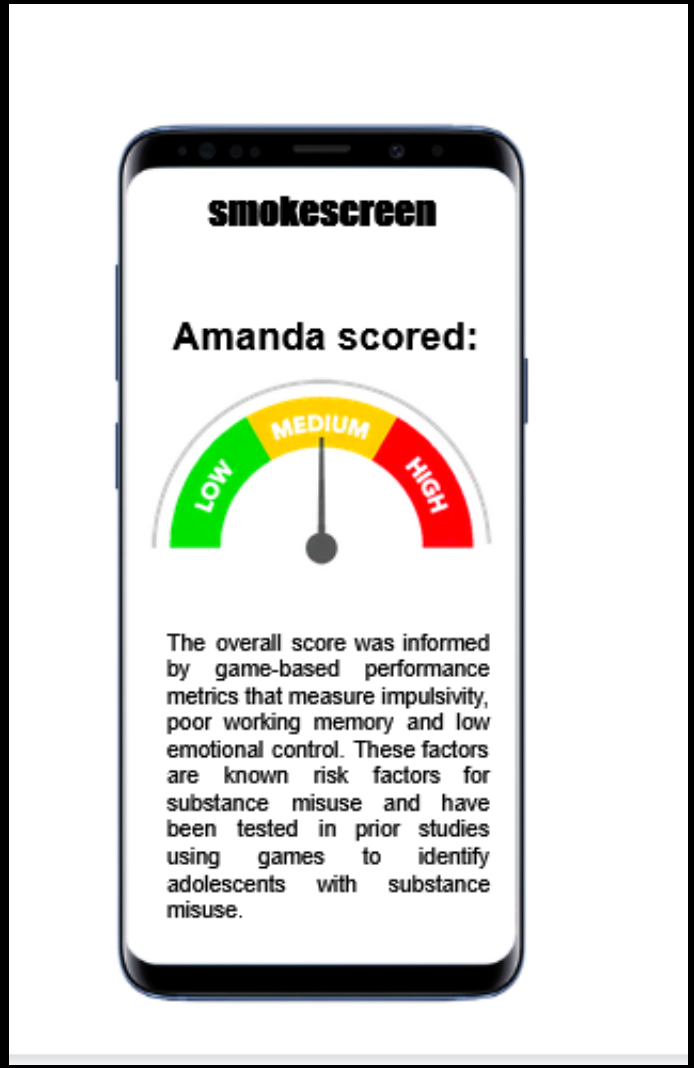
Most are willing to trade off insight into the model for greater predictive accuracy, with exceptions; they are willing to delegate information parsing to other experts.

Most stakeholders do not want to open the black box of how the AI works.

Stakeholders see value in conceptual explanations.

Substantive expert designer, SS5:

Click to learn more feature; breakdown available for those who are interested



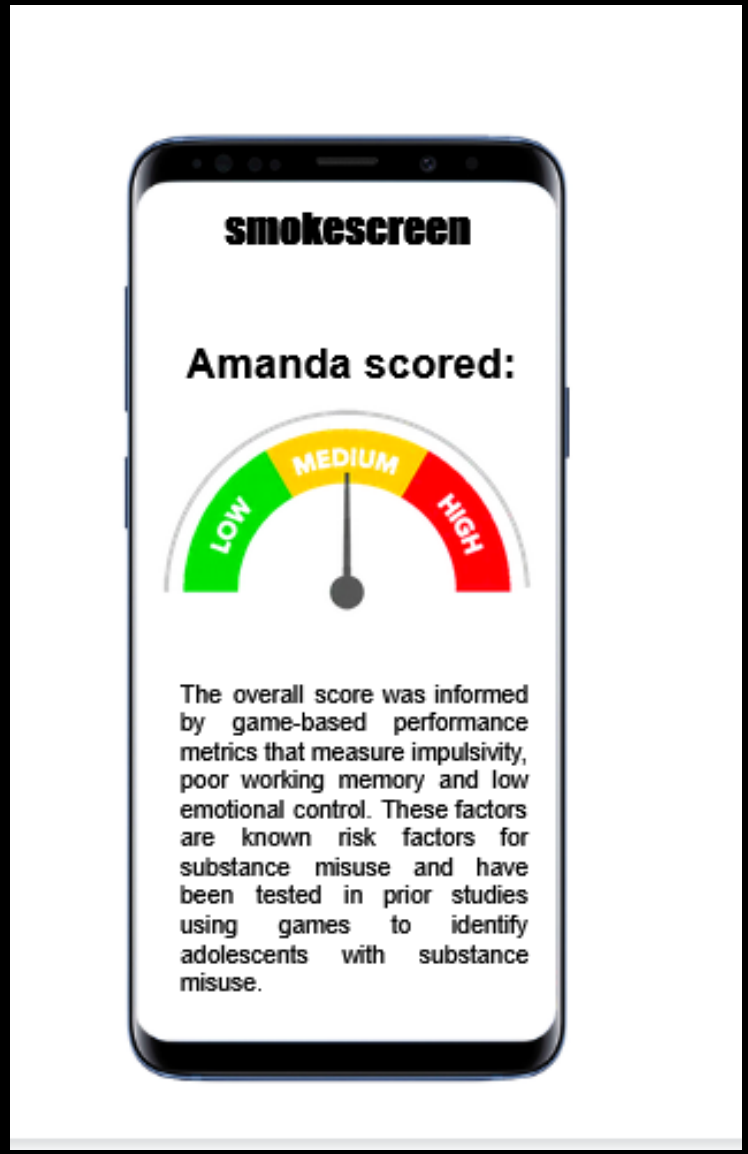
Stakeholders see value in conceptual explanations.

Teacher, SS12:

-“What about her play led to this score so that I could think about some interventions”

Making connections to risk factors for substance misuse

Suggestions about what to do next



Stakeholders are generally willing to sacrifice explainability for greater accuracy and scale.

Developer, SS10

- Trade off between accuracy and internal insight into the model and scale of reach and intervention possibilities
- Stakeholders are generally willing to sacrifice explainability for greater accuracy and scale.
- See the need for multiple explanations.

Raise the stakes and teachers think twice about black-boxed models

- Teacher SS17: “ “You can’t measure impulsivity and give [students] a grade for it in the grade book. That’s just not fair. Or emotional control. Those are the red flag indicator things that you would go talk to a counselor about.”

Concluding thoughts

- Consensus: Stakeholders have little expectation for what goes into an explanation for opaque AI
- Keep the black box closed?
- Delegate responsibility for harms to existing institutional arrangements.
- AI explanations have limited effect on legitimacy of AI for teachers, but are justifications for their use and interpretations.

Lessons learned from study 1

- Hard to repurpose previously collected data, even when clinical trial grade
  - Incomplete (compliance?), noisy, biased, non-integrated care, data is not longitudinal, selection bias in mental health clients
- Should we always use predictive AI that we can understand?
  - Inherently interpretable predictive AI models that shed insight into the “why” of the prediction offered more leverage over higher accuracy, opaque AI predictive models, even if few wanted to look inside the black box.
- AI is not incorporating our conceptual viewpoint
  - Teachers and guidance counselors want conceptual explanations largely not to believe/trust the prediction but to know what further action they might take; see value in credentialing to gain acceptance from administrators.
  - Layers of information will be necessary for implementation: Expert designers want to provide a handbook alongside explanation.

Study 2

Designing an AI predictive screening tool for diabetic retinopathy

Many predictive AI tools are just beginning to go through FDA clearance.

First autonomously approved AI prediction by FDA: Diabetic retinopathy screening

AEYE Health | Our Products | Solutions | Resources | Company | Contact us

AURORA AEYE

Cost-effective, portable retinal camera with built-in diagnostic screening intelligence

FDA CLEARED



- Portable camera enabling fast, easy and flexible screening
- Internet-connected camera automatically processes exams and provide a diagnostic result on the spot

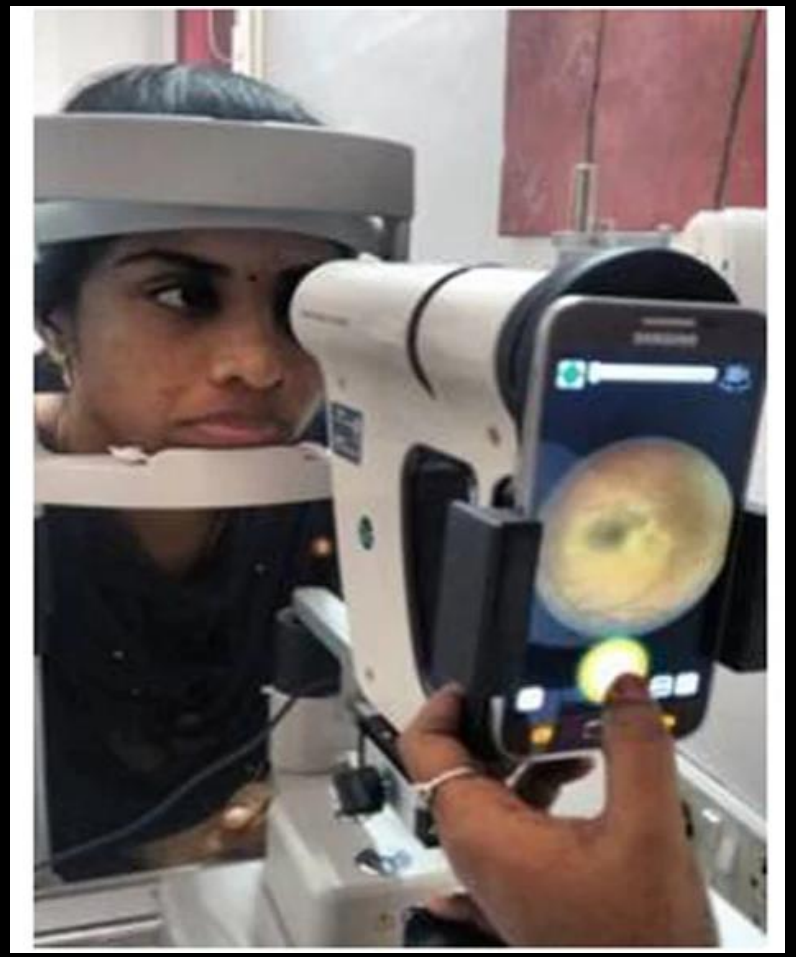
AEYE with Topcon

Robotic tabletop camera for superior ease of use



FDA CLEARED

- Robotic camera automatically captures images
- AI diagnoses images on the spot, with no need for a specialist over read



diabetic retinopathy: eye disease caused by poorly managed diabetes; primary reason for visual impairment worldwide.

AI Diabetic Retinopathy Screening

- Ethnographic observation of medical device vendors and physicians at medical conferences; interviews with experts (n=10)
- Ongoing experiments with 409 participants, master's students in a machine learning class, generating predictive AI for diabetic retinopathy (and explanations along with them)
- Asked to develop a screening tool for a smartphone usable by a community health worker in low-resource settings
  - Specifically: Provide global and individual explanations for the AI system to an audience of their choice
  - Varied conditions:
    - By purpose (none; preserving the dignity of end users; and improving human-AI collaboration)
    - By comprehensiveness (low; high)

Research Questions

- How do differences in policy requirements influence how developers design and implement end-user explanations?
- What explanations do developers design for end users?
  - How do they discuss policy compliance?
  - What aspects of designing explanations for end users do they find challenging?

Findings overview

- Based on prior research, we expected that providing a purpose and more comprehensive policy guidance would lead to higher quality explanations to contextualize the AI prediction.
  - For global, expert explanations if not for local, individual explanations for predictions
- We found that **participants across the board struggled to produce quality explanations**, comply with the provided policy requirements for explainability, and provide evidence of compliance.
- We posit that participant noncompliance is in part due to a failure to imagine and anticipate the needs of their audience, particularly non-technical stakeholders.

Toward better end user explanations

1) Lack of technical understanding of explainability tools

2) Explanations for developers' needs and expertise

3) Explanations that failed to differentiate between end users

4) Explanations for end users without embedding in system context and purpose

Good explanations: Embedded in system context

2. Prediction Results

Diabetic Retinopathy Severity Level: Mild

Confidence Scores:

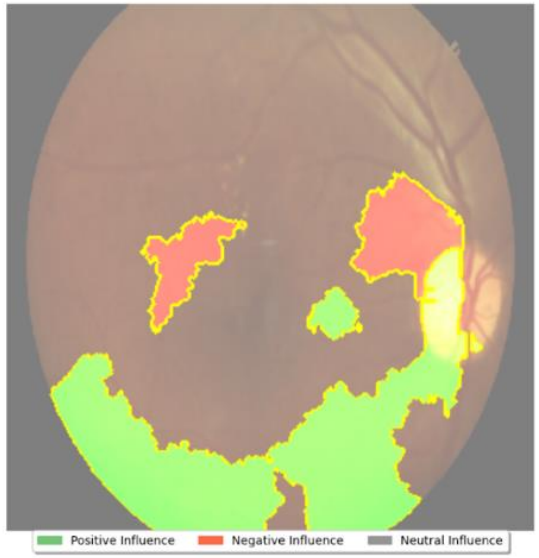
- No DR: 0.48%
- Mild: 91.42%
- Moderate: 7.94%
- Severe: 0.02%
- Proliferative DR: 0.14%

The confidence scores presented here reflect the likelihood of each diabetic retinopathy severity level based on the model's analysis.

Note: The above percentages are generated by a machine learning model and represent probabilistic assessments, not certainties. These results should be used as one of several tools in a comprehensive diagnostic process and not relied upon solely for medical decisions.

3. Diagnostic Insights

Highlighted Features Influencing Diagnosis Using LIME:



4. Understanding LIME Visualizations

LIME (Local Interpretable Model-Agnostic Explanations) explains predictions of machine learning models by approximating them locally with an interpretable model. It identifies and highlights the regions in the input image that significantly impact the model's output. Learn more about LIME on its official LIME.

- **Positive Influence:** Green indicates that certain regions or pixels in the image have a significant contribution towards the model's prediction of a particular class.
- **Negative Influence:** Red indicates that certain regions or pixels in the image have a significant contribution against the model's prediction of a particular class.
- **Neutral Influence:** Grey Areas that do not significantly impact the prediction's outcome.

5. Interpretation and Next Steps

Based on the severity level and highlighted features, consider the following actions:

- If 'Mild' or 'No DR': Schedule annual screening.
- If 'Moderate': Refer to an ophthalmologist for possible intervention.
- If 'Severe' or 'Proliferative DR': Urgent referral to a specialist is recommended.

Beyond diabetic retinopathy

- Respondent 3, sales manager, medical tech company
- “The AI models that are currently out there that’re like accessible to most clinicians that I'm aware of, are only looking for diabetic retinopathy. There's about 311 other disease states that we can find with a retina screening, and the AI models do not comment on any of those… Our primary market are looking for the other 311 of things too.”
- Explanations and predictive ability still don’t quite meet healthcare expectations.
- Frontline software engineers lack the toolkit for thinking about transparency and systems and need more support.

Key Takeaways for Study 2

- Since 2024, use of diabetic retinopathy screening tools have skyrocketed, aided by an autonomous AI insurance reimbursement code.
  - Adoption has been more successful in larger health systems, despite significant pushback
  - Adoption facilitated by the inclusion of diabetic retinopathy screening as a quality metric for clinics (e.g., MIPS or HEDIS)
  - Viability and parallels in the space of behavioral health?
- AI formalizes clinic practices around historical precedents rather than leading to new, streamlined workflows.

Takeaways for Predictive AI and Behavioral Health

- There are significant growing pains in predictive AI for health.
- Basic transparency disclosing when an AI system is in use is not the norm (but our previous study shows, matters to stakeholders).
- There are many considerations for predictive AI tools with respect to ethics or responsible AI in addition to the important headline concerns about fairness and equity: Transparency and explainability as important.
- The general public does not know what to expect of AI, and the baseline for developers about what information to share about an AI system is limited.
  - Particularly important in health contexts
  - People as a part of the equation



Thank you!

- This work was supported by a National Science Foundation Grant, CISE/CCF 2131466.
- Thanks to research assistants: Bridgitte Thao, Matthew Bray, Fikir Mekonnen, Ines Gilles, Jacob Tjaden, Jenny Rowlett, Nadia Nahar, Jane Hsieh, Shirin Purkayastha and Zahra Abba Omar

- Email: alka.menon@yale.edu

Next steps

- We have run this experiment several times since fall 2024, varying additional conditions and providing more guidance:
  - Testing the effect of scoping the policy conditions (giving even more guidance)
  - Testing the effect of requiring chatbot interaction to simulate sociological imagination

Explanation Problem 1: Nonsense

Diagnosis: Moderate

This image is predicted to belong to class 2 with confidence 0.86. Shap_Summary represents a summary of the SHAP (SHapley Additive exPlanations) analysis results. It provides a concise overview of the SHAP values calculated for each feature in the input data, typically indicating the impact of each feature on the model's prediction. SHAP analysis summary: 0: 0.00, 1: 0.00, 2: 0.00, 3: 0.00, 4: 0.00, 5: 0.00, 6: 0.00, 7: 0.00, 8: 0.00, 9: 0.00, 10: 0.00, 11: 0.00, 12: 0.00, 13: 0.00, 14: 0.00, 15: 0.00, 16: 0.00, 17: 0.00, 18: 0.00, 19: 0.00, 20: 0.00, 21: 0.00, 22: 0.00, 23: 0.00, 24: 0.00, 25: 0.00, 26: 0.00, 27: 0.00, 28: 0.00, 29: 0.00, 30: 0.00, 31: 0.00, 32: 0.00, 33: 0.00, 34: 0.00, 35: 0.00, 36: 0.00, 37: 0.00, 38: 0.00, 39: 0.00, 40: 0.00, 41: 0.00, 42: 0.00, 43: 0.00, 44: 0.00, 45: 0.00, 46: 0.00, 47: 0.00, 48: 0.00, 49: 0.00, 50: 0.00, 51: 0.00, 52: 0.00, 53: 0.00, 54: 0.00, 55: 0.00, 56: 0.00, 57: 0.00, 58: 0.00, 59: 0.00, 60: 0.00, 61: 0.00, 62: 0.00, 63: 0.00, 64: 0.00, 65: 0.00, 66: 0.00, 67: 0.00, 68: 0.00, 69: 0.00, 70: 0.00, 71: 0.00, 72: 0.00, 73: 0.00, 74: 0.00, 75: 0.00, 76: 0.00, 77: 0.00, 78: 0.00, 79: 0.00, 80: 0.00, 81: 0.00, 82: 0.00, 83: 0.00, 84: 0.00, 85: 0.00, 86: 0.00, 87: 0.00, 88: 0.00, 89: 0.00, 90: 0.00, 91: 0.00, 92: 0.00, 93: 0.00, 94: 0.00, 95: 0.00, 96: 0.00, 97: 0.00, 98: 0.00, 99: 0.00, 100: 0.00, 101: 0.00, 102: 0.00, 103: 0.00, 104: 0.00, 105: 0.00, 106: 0.00, 107: 0.00, 108: 0.00, 109: 0.00, 110: 0.00, 111: 0.00, 112: 0.00, 113: 0.00, 114: 0.00, 115: 0.00, 116: 0.00, 117: 0.00, 118: 0.00, 119: 0.00, 120: 0.00, 121: 0.00, 122: 0.00, 123: 0.00, 124: 0.00, 125: 0.00, 126: 0.00, 127: 0.00, 128: 0.00, 129: 0.00, 130: 0.00, 131: 0.00, 132: 0.00, 133: 0.00, 134: 0.00, 135: 0.00, 136: 0.00, 137: 0.00, 138: 0.00, 139: 0.00, 140: 0.00, 141: 0.00, 142: 0.00, 143: 0.00, 144: 0.00, 145: 0.00, 146: 0.00, 147: 0.00, 148: 0.00, 149: 0.00, 150: 0.00, 151: 0.00, 152: 0.00, 153: 0.00, 154: 0.00, 155: 0.00, 156: 0.00, 157: 0.00, 158: 0.00, 159: 0.00, 160: 0.00, 161: 0.00, 162: 0.00, 163: 0.00, 164: 0.00, 165: 0.00, 166: 0.00, 167: 0.00, 168: 0.00, 169: 0.00, 170: 0.00, 171: 0.00, 172: 0.00, 173: 0.00, 174: 0.00, 175: 0.00, 176: 0.00, 177: 0.00, 178: 0.00, 179: 0.00, 180: 0.00, 181: 0.00, 182: 0.00, 183: 0.00, 184: 0.00, 185: 0.00, 186: 0.00, 187: 0.00, 188: 0.00, 189: 0.00, 190: 0.00, 191: 0.00, 192: 0.00, 193: 0.00, 194: 0.00, 195: 0.00, 196: 0.00, 197: 0.00, 198: 0.00, 199: 0.00, 200: 0.00, 201: 0.00, 202: 0.00, 203: 0.00, 204: 0.00, 205: 0.00, 206: 0.00, 207: 0.00, 208: 0.00, 209: 0.00, 210: 0.00, 211: 0.00, 212:

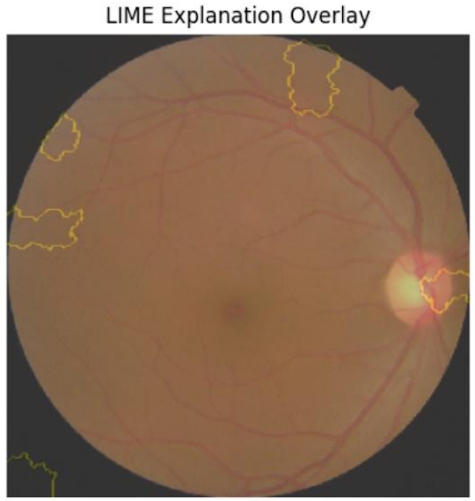
Explanation Problem 2: For developer needs

- Prediction explanation patient 03
- Age: 79
- Sex: F
- Diagnosis: No DR
- This image highlights important features for analysis.



Explanation Problem 3: Too medical jargon

Local Interpretable Model-agnostic Explanations: LIME



For the user: LIME generates localized explanations by highlighting the most important regions of the input image that influence the model's prediction. LIME indicates specific areas of the eye image that contribute most significantly to the model's prediction of retinopathy progression. Interpreting these regions should be left to a medical professional.

For the medical professional: LIME provides insights into the specific features or abnormalities in the retinal image that drive the model's predictions. Medical professionals can use this information to validate the model's decisions and assess whether the highlighted regions align with known indicators of retinopathy progression. Additionally, LIME explanations can help medical professionals identify areas of concern in the retinal image that may require further examination or intervention.