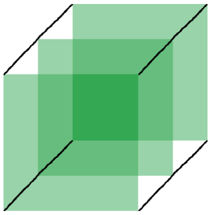
Explainability + Trust

Chapter worksheet



Instructions

Block out time to get as many cross-functional leads as possible together in a room to work through these exercises & checklists.

Exercises

1. Trust calibration [~1 hour]

Imagine situations where users could under-trust or over-trust your feature.

2. Explanation strategy[~30 minutes]

Determine which user interactions require an explanation, and what kind.

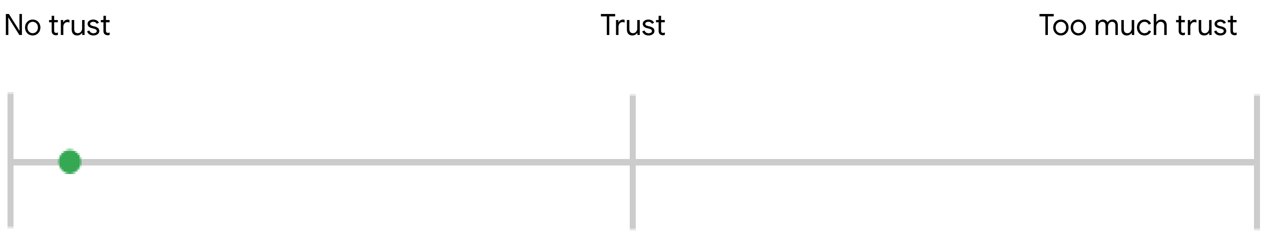
3. Test with user research [multiple sessions]

Make sure that the explanation visuals and messaging make sense and are helpful for your users.

1. Trust calibration

As a team, brainstorm what kinds of experiences and interactions would decrease, maintain, or inflate trust in your feature's AI. Identify the underlying data sources, system data and user knowledge, that could impact the calibration.

Example product: AI that classifies a skin condition.



No trust

Trust

Too much trust

User Group A

Example user group: Doctors of patients using the AI system

Example scenario: Patients see doctor with a condition that was mis-classified by the AI.

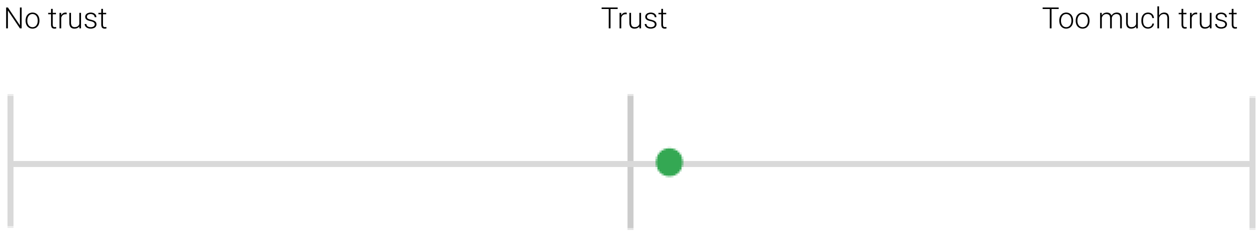
System data impacting calibration	*Example: image of current condition submitted by user, label of skin conditions in training data*
User knowledge impacting calibration	*Example: user's prior medical history*

User Group B

Example user group: Patients using the AI system

Example scenario: Patients see doctor with a condition that was misclassified by the AI.

System data impacting calibration	
User knowledge impacting calibration	



User Group A

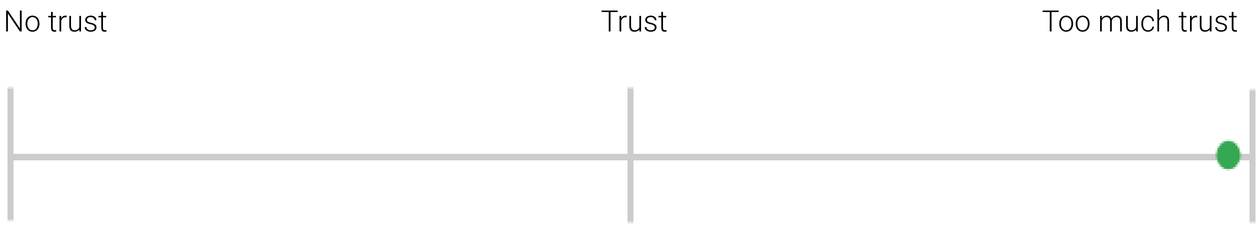
Example user group: Patients using the AI system

Example scenario: Patient uses AI system to identify a common and temporary condition, like poison ivy, and receives a recommended treatment that works.

System data impacting calibration	
User knowledge impacting calibration	

User Group B

System data impacting calibration	
User knowledge impacting calibration	



User Group A

Example user group: Patients using the AI system

Example scenario: Patients with pre-cancerous cells doesn't double check the app's diagnosis

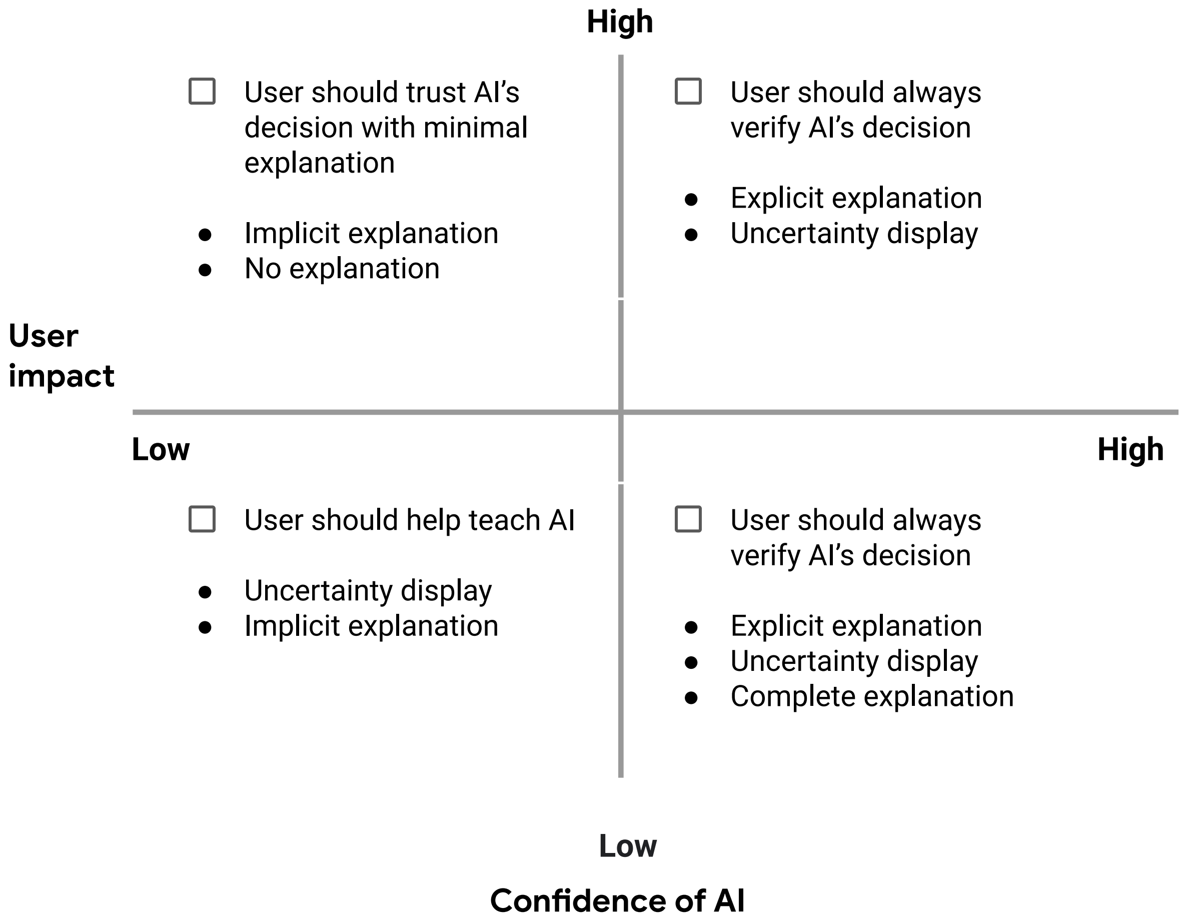
System data impacting calibration	
User knowledge impacting calibration	

User Group B

System data impacting calibration	
User knowledge impacting calibration	

2. Explanation strategy

After mapping the range of interactions that could decrease, maintain, or inflate trust, decide which interactions require an explanation. Use the 2x2 template of “User impact” vs. “Confidence of AI” to help narrow down which explanations you want to test with users. Once you have consensus on the interactions that require an explanation, use the templates below to draft the explanation copy for user testing.



High

□ User should trust AI’s decision with minimal explanation

- Implicit explanation
- No explanation

□ User should always verify AI’s decision

- Explicit explanation
- Uncertainty display

User impact

Low

High

□ User should help teach AI

- Uncertainty display
- Implicit explanation

□ User should always verify AI’s decision

- Explicit explanation
- Uncertainty display
- Complete explanation

Low

Confidence of AI

Explanation messaging

Product interaction	
Users	
Technical understanding	☐ expert ☐ non-expert
Explanation type	
Explanation text	

Messaging templates

Explicit explanation

Example:
“You are seeing this video recommendation because you often watch cooking videos.”
“This is most likely ____X_____, because _____Y______.”

Uncertainty display - Confidence level

Example: “Prediction: [category] XX%”

Uncertainty display - N-Best

Example: “Most likely X, Y, or Z”

3. Test with user research

Use the explanations you have drafted in user research. Multiple research efforts will be needed to understand any trust concerns users anticipate, and arrive at the ideal explanations for your user groups.

Research protocol questions

- On this scale, show me how trusting you are of this recommendation. [show scale]
- What questions do you have about how [name of product / feature] came to this recommendation?
- What, if anything, would increase your trust in this recommendation?
- How satisfied or dissatisfied are you with the explanation written here?

Once you're through with the above, ask additional questions to gauge user understanding of your specific AI system and the actions users can take.