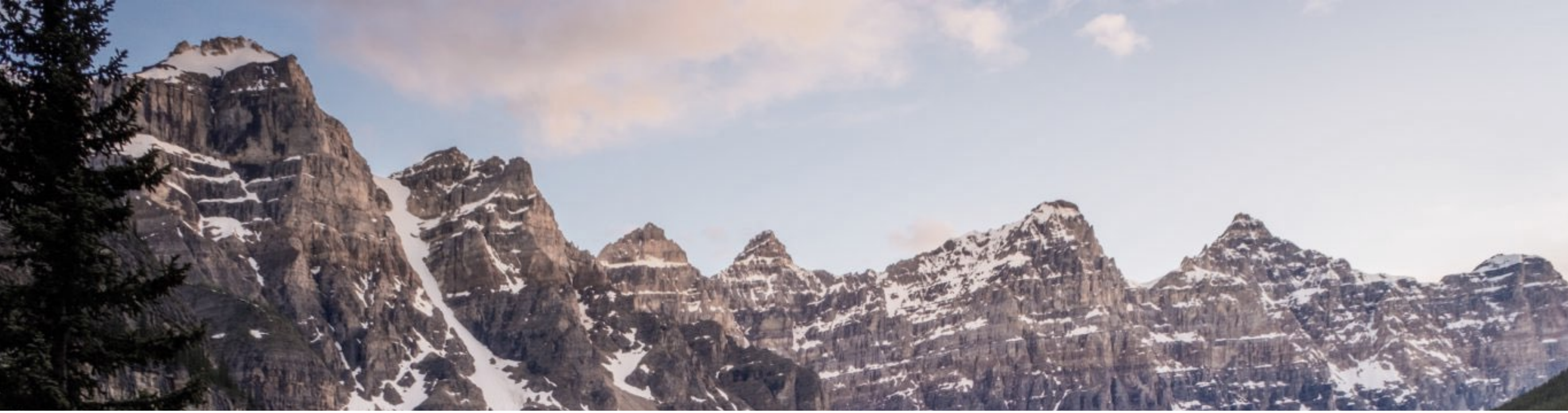
A Psychological Theory of Explainability

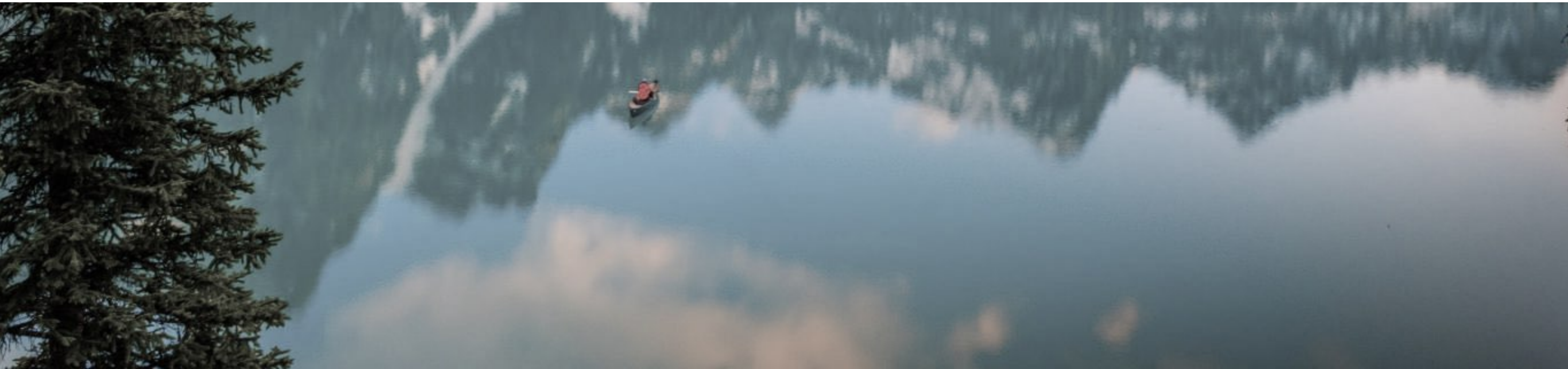
Scott Cheng-Hsin Yang*, Tomas Folke* & Patrick Shafto
*equal contribution

The goal of eXplainable Artificial Intelligence (XAI) is to make AI decision **understandable to humans.**

- MANY techniques to generate explanations
- Analysis of the techniques
- Validation of the techniques
- How humans interpret the explanations given



Humans **project their beliefs** onto the AI; they interpret the explanation provided by **comparing it to the explanations that they themselves would give.**



Machine faithfulness

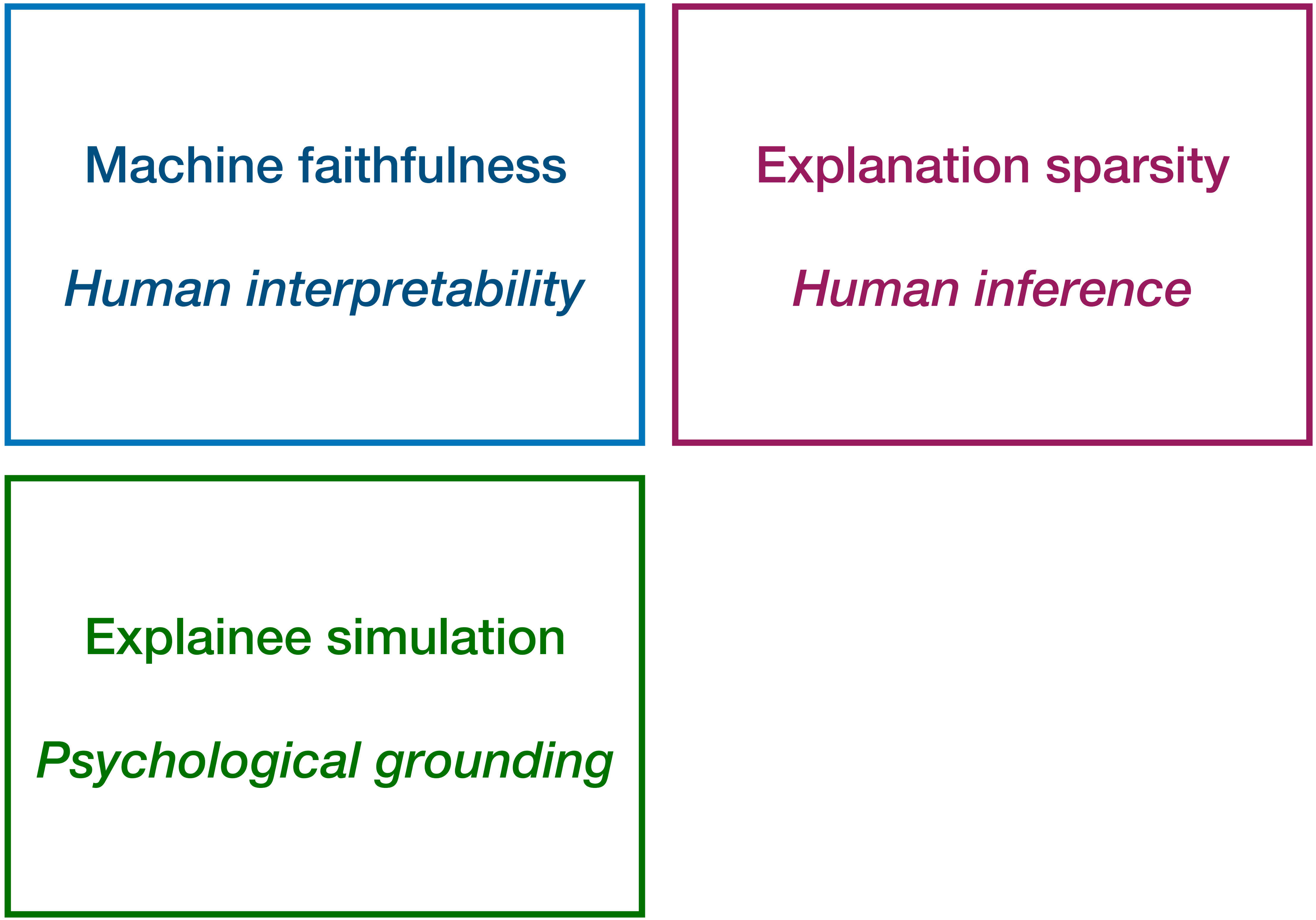
Human interpretability

Machine faithfulness

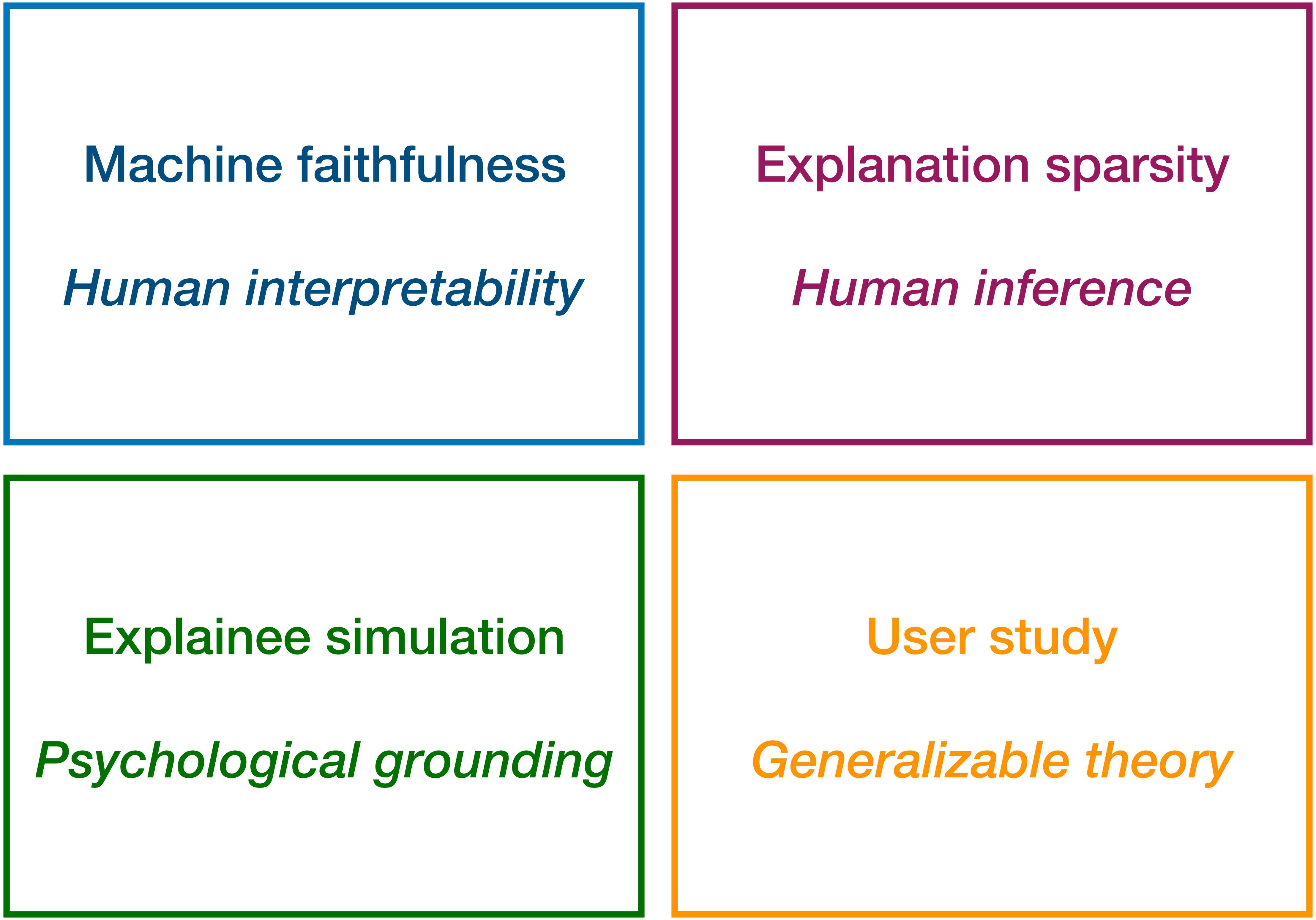
Human interpretability

Explanation sparsity

Human inference



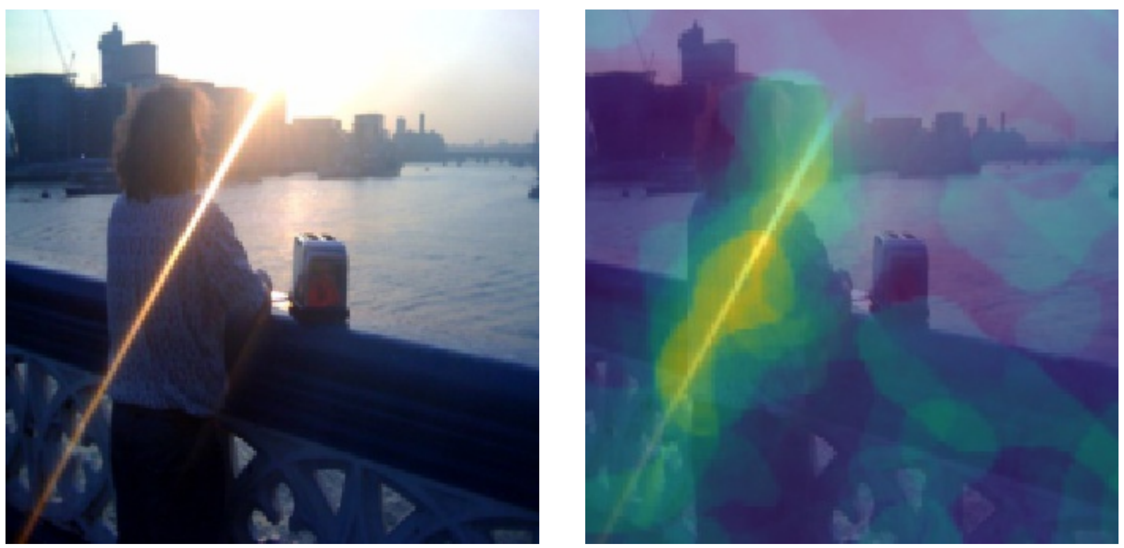
Machine faithfulness
Human interpretability
Explanation sparsity
Human inference
Explainee simulation
Psychological grounding



Machine faithfulness
Human interpretability
Explanation sparsity
Human inference
Explainee simulation
Psychological grounding
User study
Generalizable theory

Example trial
(Explanation condition)

Which category do you think the robot will classify the image as?



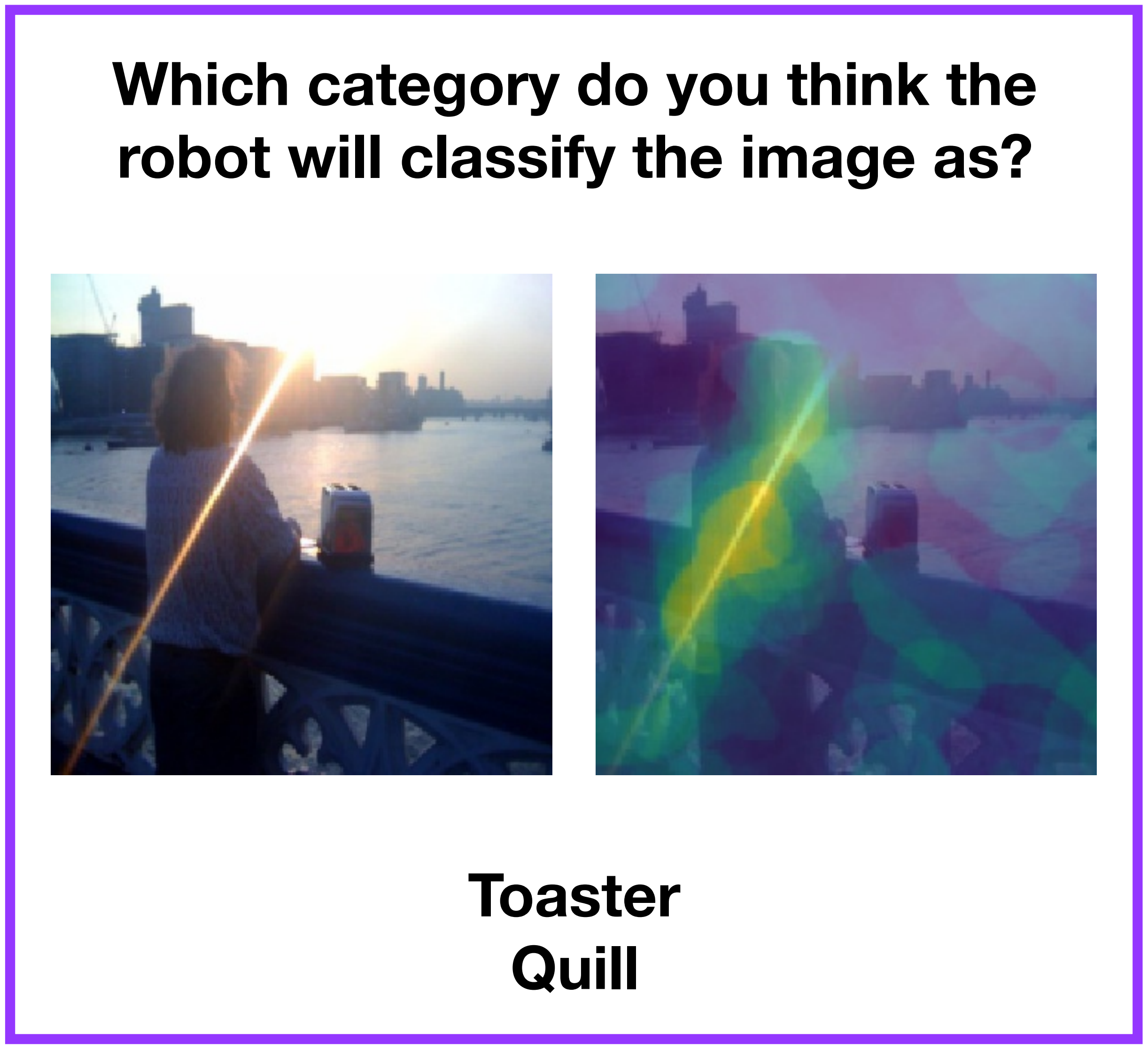
Toaster
Quill

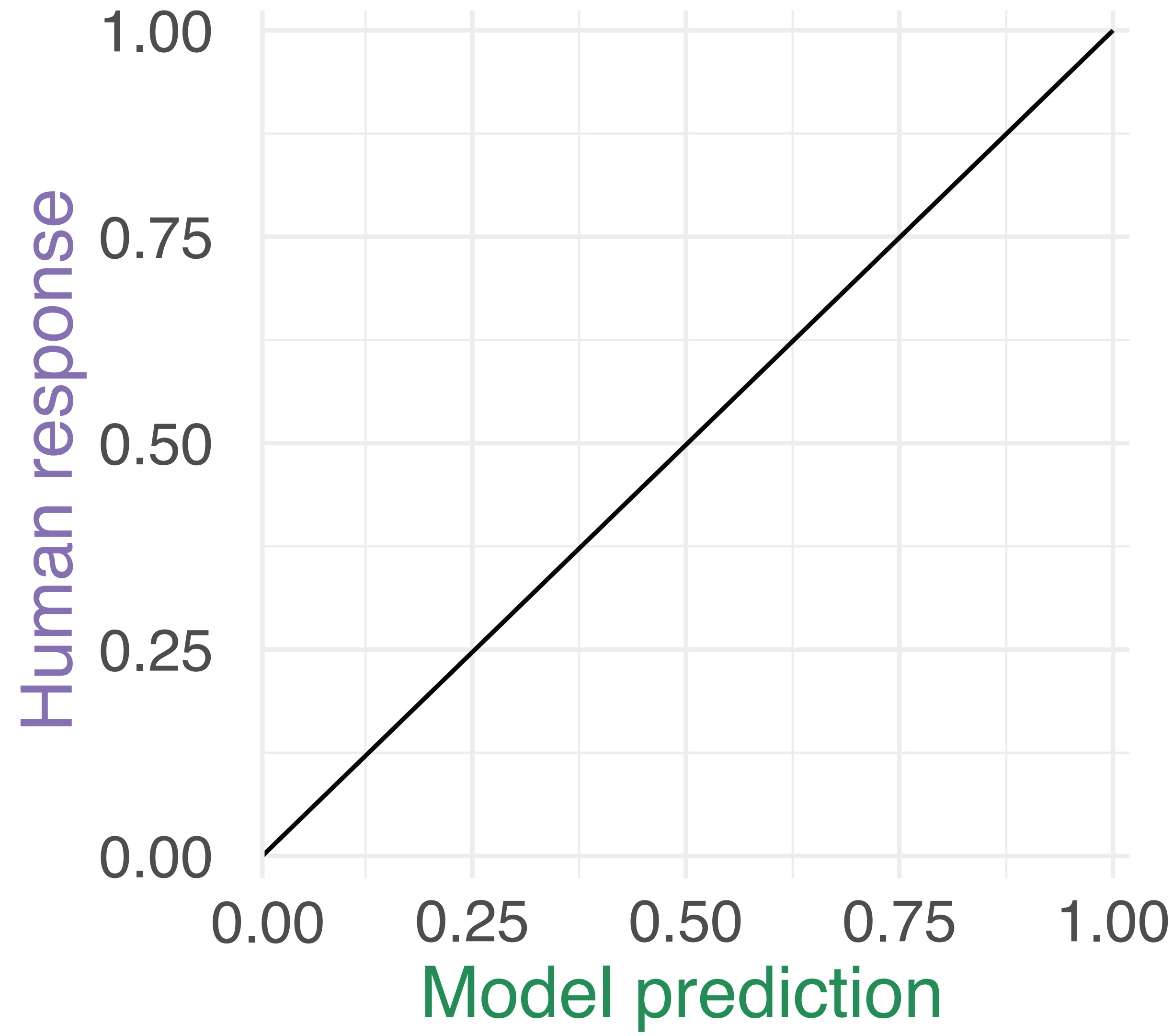
AI to be explained: ResNet-50 trained on ImageNet

Explanation: Saliency maps generated from Bayesian Teaching

Task: predict AI classification

Example trial
(Explanation condition)





$$\overset{\text{Posterior}}{\boxed{P(c \mid \mathbf{e}, \mathbf{x})}} \propto \overset{\text{Prior}}{\boxed{P(c \mid \mathbf{x})}} \overset{\text{Likelihood}}{\boxed{p(\mathbf{e} \mid c, \mathbf{x})}}$$

Theory's prediction of
human response

$$\overset{\textbf{Posterior}}{\boxed{P(c \mid \mathbf{e}, \mathbf{x})}} \propto \overset{\textbf{Prior}}{\boxed{P(c \mid \mathbf{x})}} \overset{\textbf{Likelihood}}{\boxed{p(\mathbf{e} \mid c, \mathbf{x})}}$$

AI classification

explanation

image

Theory's prediction of human response

Human assumption about the AI

Posterior **Prior** **Likelihood**

$$\boxed{P(c \mid \mathbf{e}, \mathbf{x})} \propto \boxed{P(c \mid \mathbf{x})} \boxed{p(\mathbf{e} \mid c, \mathbf{x})}$$

AI classification

explanation

image

Effect of the explanation

Theory's prediction of human response

Human assumption about the AI

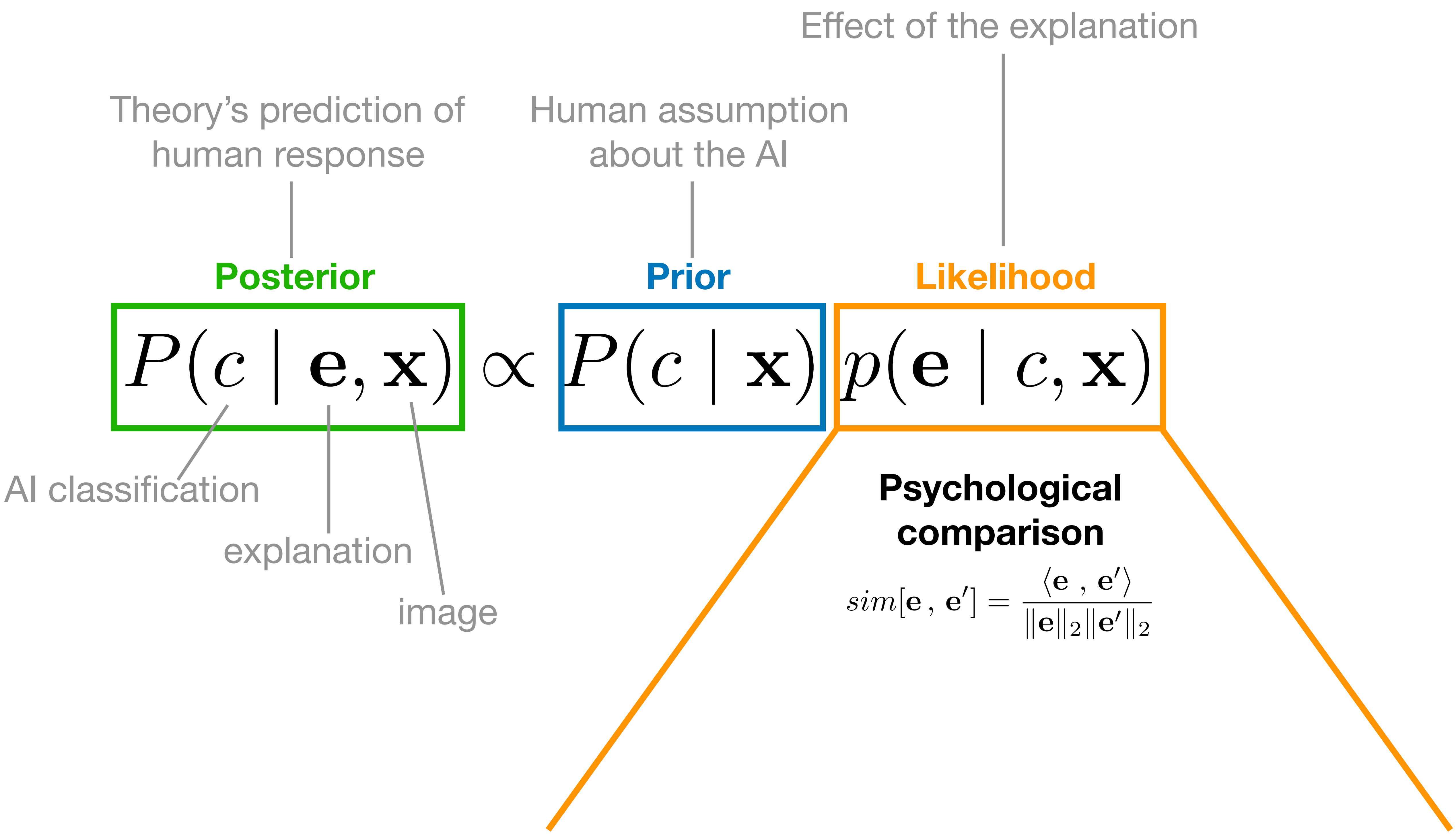
Posterior **Prior** **Likelihood**

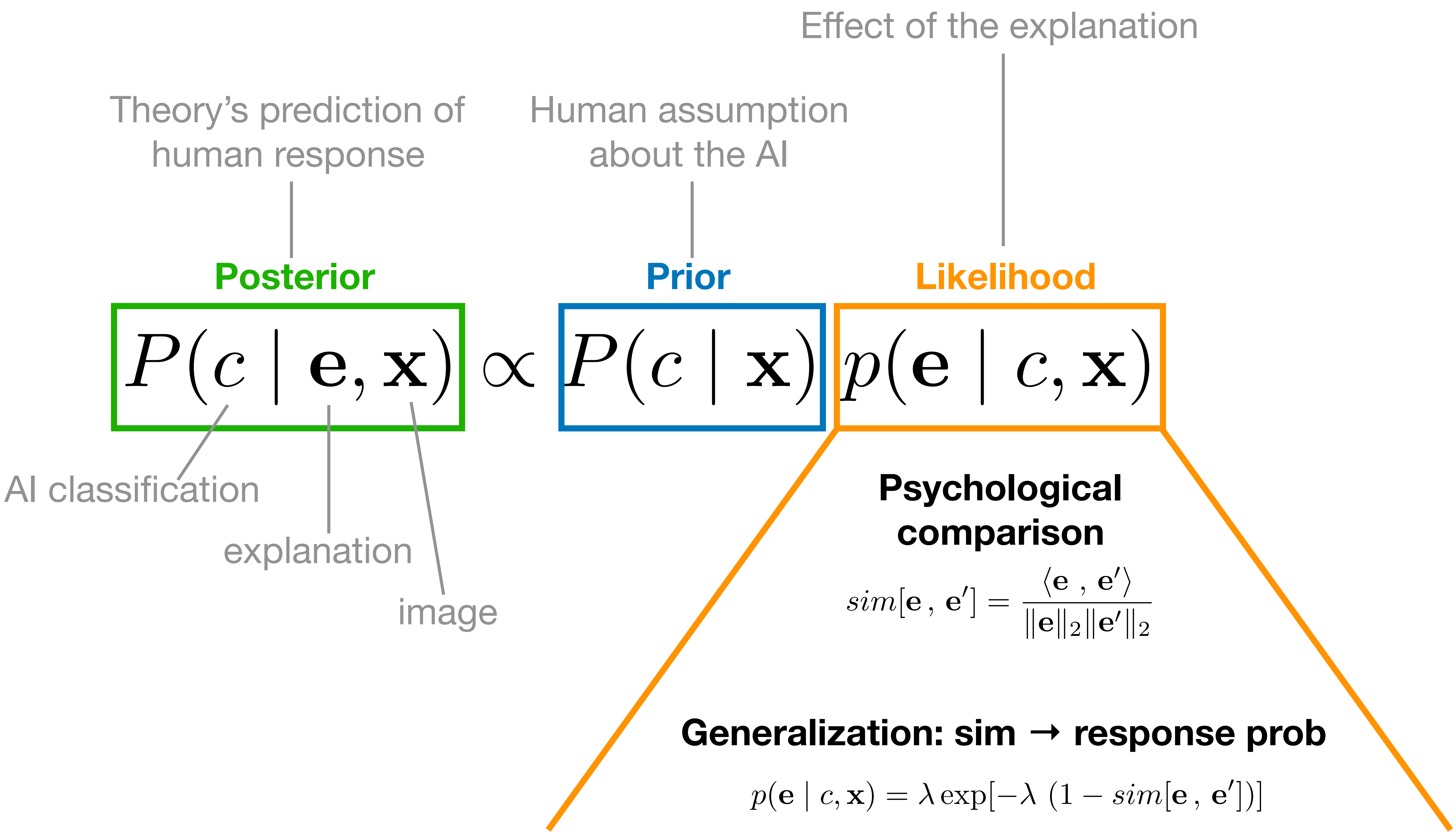
$$\boxed{P(c \mid \mathbf{e}, \mathbf{x})} \propto \boxed{P(c \mid \mathbf{x})}\,\boxed{p(\mathbf{e} \mid c, \mathbf{x})}$$

AI classification

explanation

image

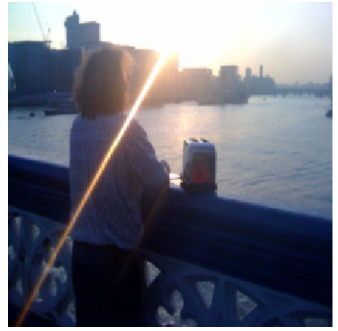

Effect of the explanation
Theory's prediction of human response
Human assumption about the AI
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
AI classification
explanation
image
Psychological comparison
$sim[\mathbf{e}\,,\,\mathbf{e}'] = \frac{\langle \mathbf{e}\,,\,\mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$



Effect of the explanation
Theory's prediction of human response
Human assumption about the AI
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
AI classification
explanation
image
Psychological comparison
$sim[\mathbf{e}\,,\,\mathbf{e}'] = \frac{\langle \mathbf{e}\,,\,\mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$
Generalization: sim → response prob
$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp[-\lambda\ (1 - sim[\mathbf{e}\,,\,\mathbf{e}'])]$

$$\overset{\text{Posterior}}{\boxed{P(c \mid \mathbf{e}, \mathbf{x})}} \propto \overset{\text{Prior}}{\boxed{P(c \mid \mathbf{x})}} \, \overset{\text{Likelihood}}{\boxed{p(\mathbf{e} \mid c, \mathbf{x})}}$$

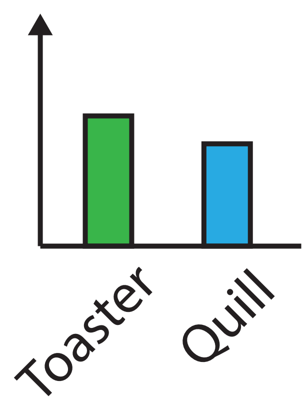
Which category do you think the robot will classify the image as?



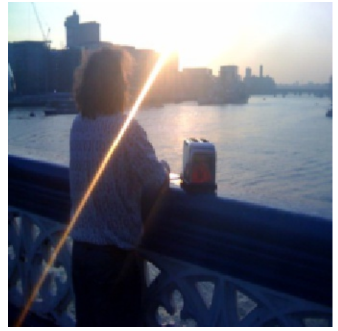
Toaster
Quill

Posterior **Prior** **Likelihood**

$$\boxed{P(c \mid \mathbf{e}, \mathbf{x})} \propto \boxed{P(c \mid \mathbf{x})}\,\boxed{p(\mathbf{e} \mid c, \mathbf{x})}$$

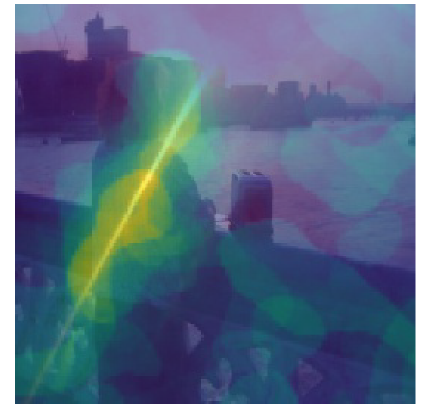


Which category do you think the robot will classify the image as?



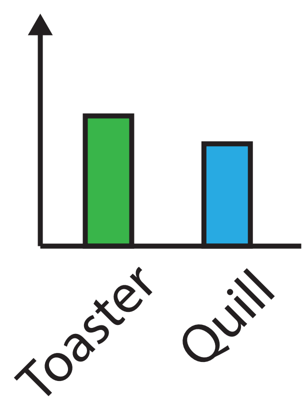
Toaster
Quill

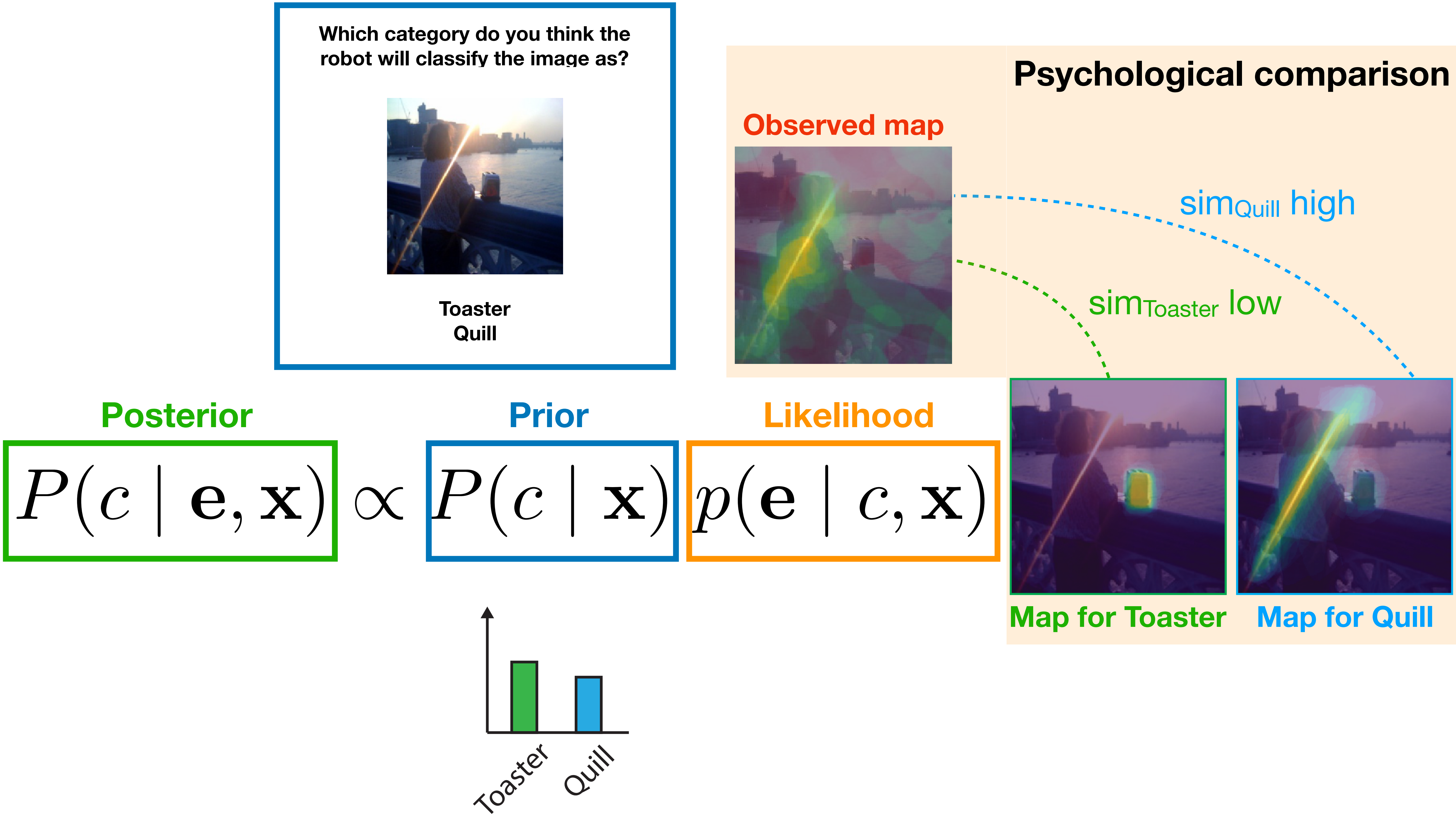
Observed map



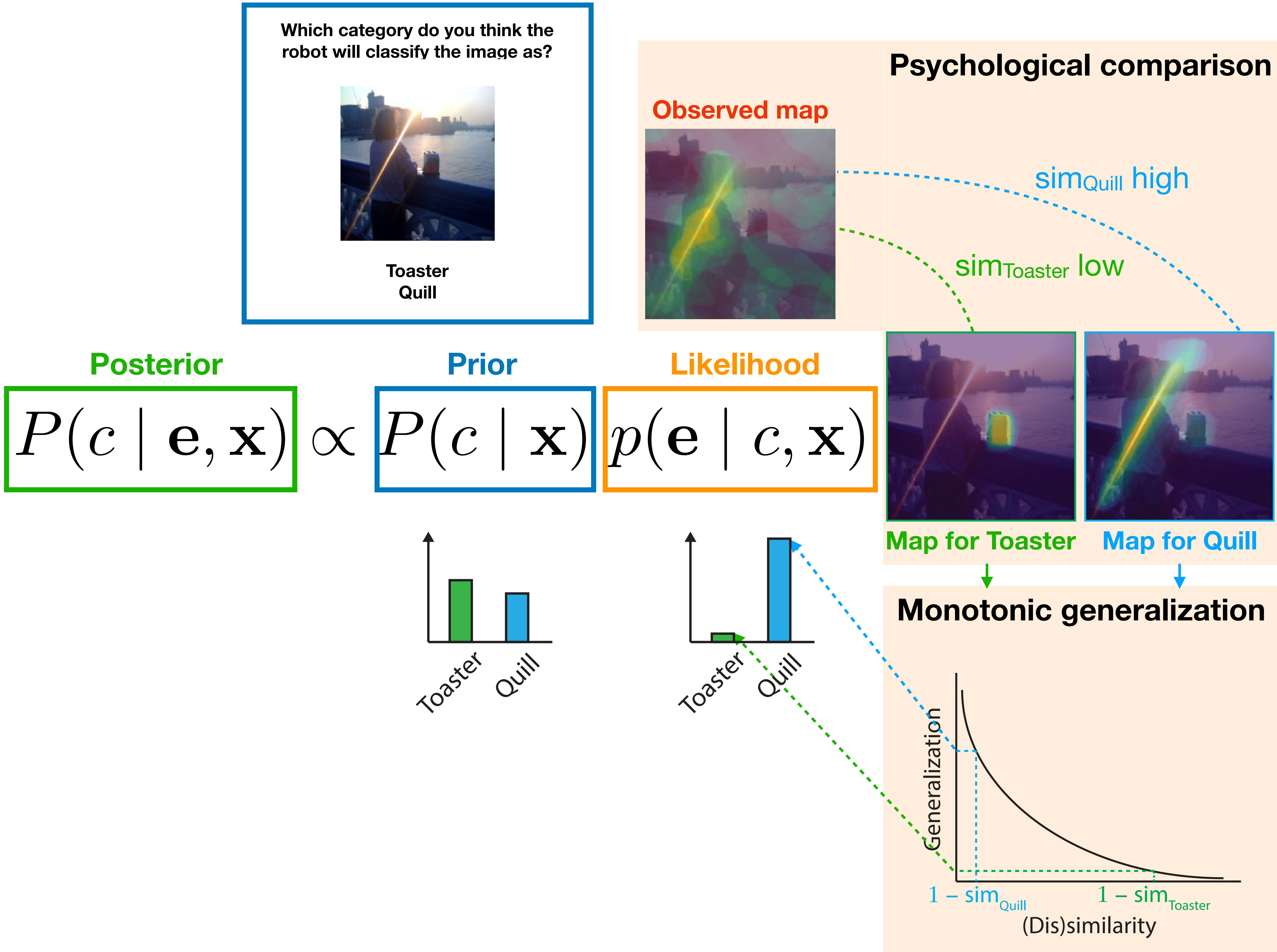
Posterior | Prior | Likelihood

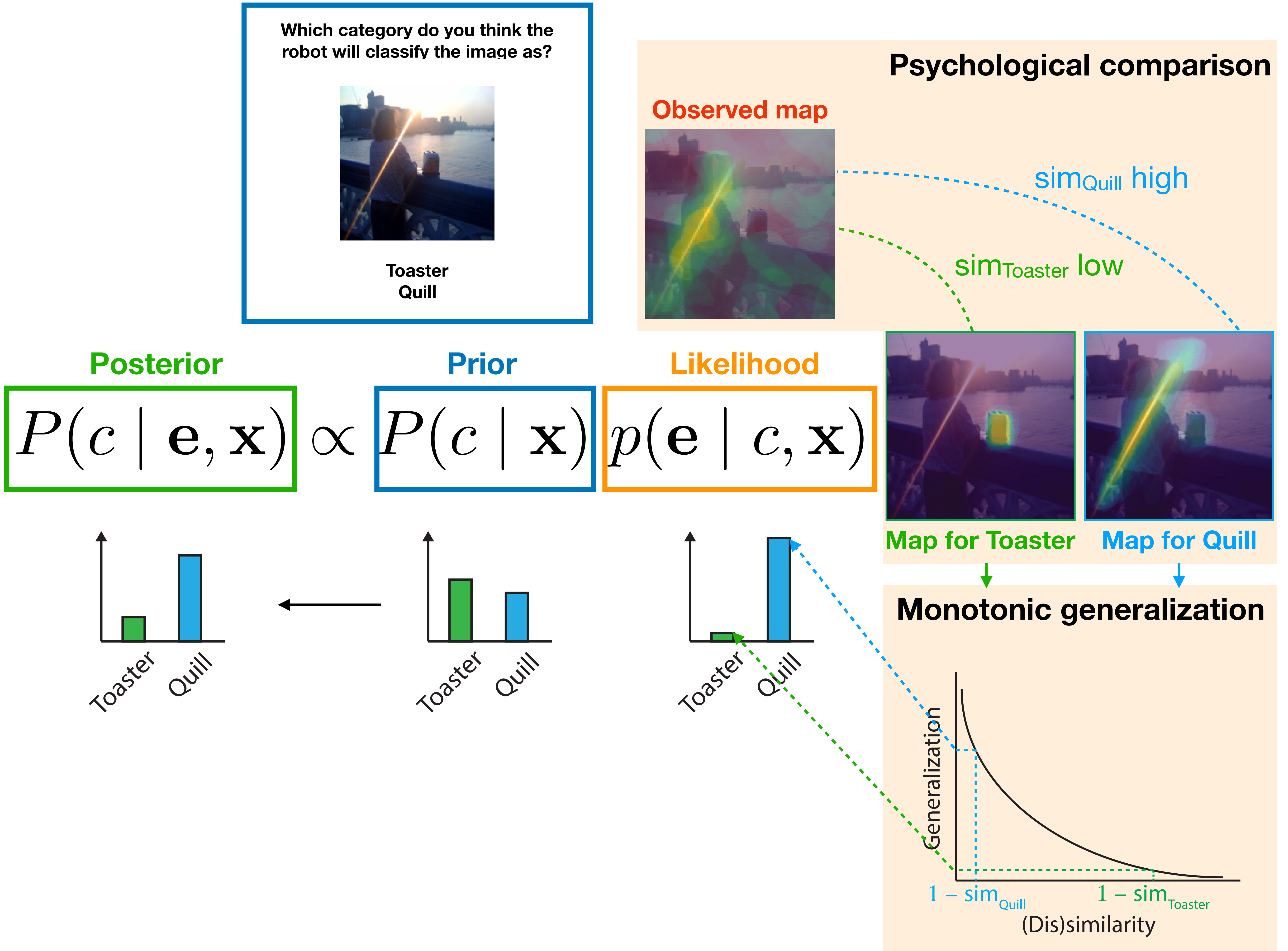
$$\boxed{P(c \mid \mathbf{e}, \mathbf{x})} \propto \boxed{P(c \mid \mathbf{x})}\,\boxed{p(\mathbf{e} \mid c, \mathbf{x})}$$





Which category do you think the robot will classify the image as?
Toaster
Quill
Psychological comparison
Observed map
sim_{Quill} high
sim_{Toaster} low
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Map for Toaster
Map for Quill
Toaster
Quill

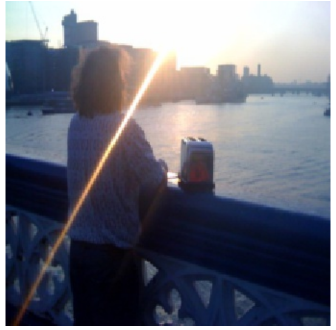

Which category do you think the robot will classify the image as?
Toaster
Quill
Psychological comparison
Observed map
sim$_{Quill}$ high
sim$_{Toaster}$ low
Map for Toaster
Map for Quill
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Toaster
Quill
Toaster
Quill
Monotonic generalization
Generalization
1 – sim$_{Quill}$
1 – sim$_{Toaster}$
(Dis)similarity



Which category do you think the robot will classify the image as?
Toaster
Quill
Psychological comparison
Observed map
sim_Quill high
sim_Toaster low
Map for Toaster
Map for Quill
Monotonic generalization
Generalization
1 – sim_Quill
1 – sim_Toaster
(Dis)similarity
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Toaster
Quill
Toaster
Quill
Toaster
Quill

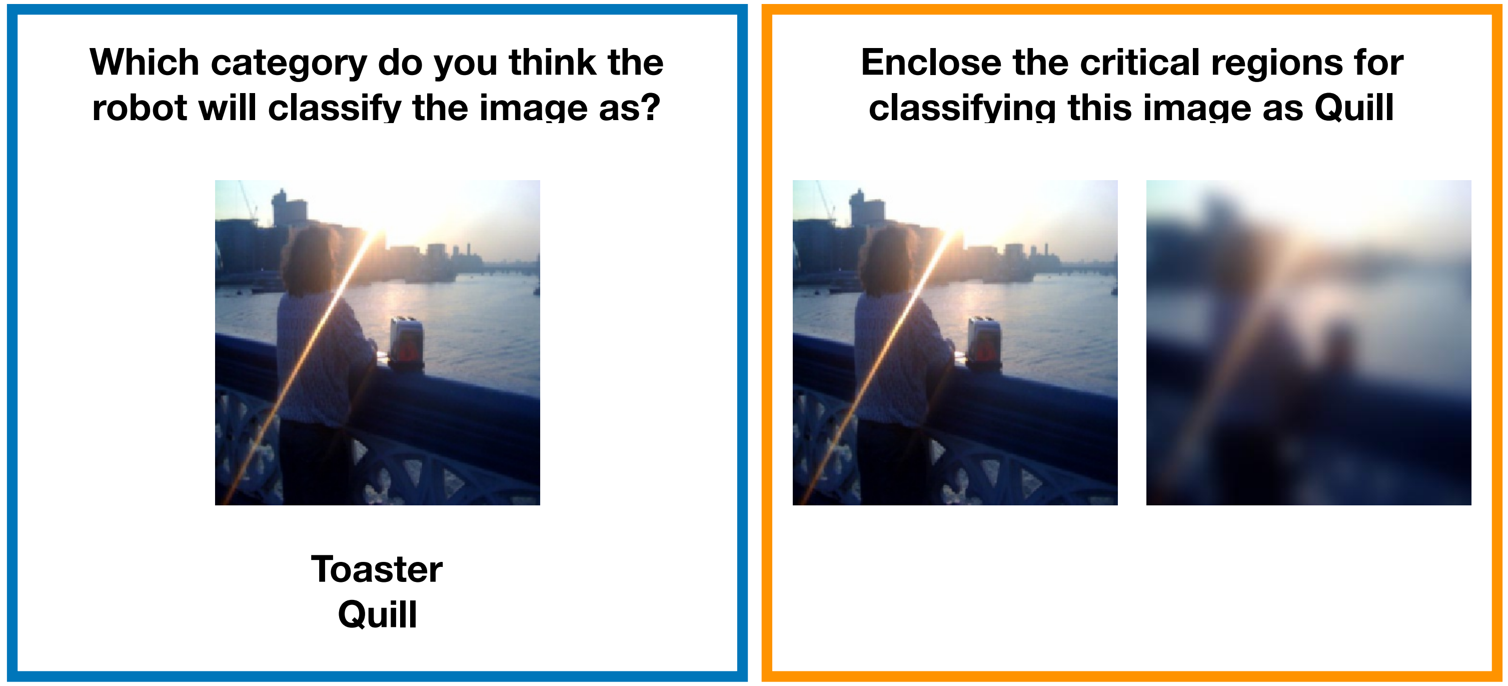
$$\underbrace{\boxed{P(c \mid \mathbf{e}, \mathbf{x})}}_{\text{Posterior}} \propto \underbrace{\boxed{P(c \mid \mathbf{x})}}_{\text{Prior}} \underbrace{\boxed{p(\mathbf{e} \mid c, \mathbf{x})}}_{\text{Likelihood}}$$

Which category you think the robot will classify the image as?



Toaster
Quill

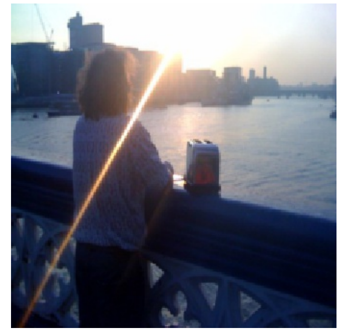
$$\overset{\text{Posterior}}{\boxed{P(c \mid \mathbf{e}, \mathbf{x})}} \propto \overset{\text{Prior}}{\boxed{P(c \mid \mathbf{x})}}\,\overset{\text{Likelihood}}{\boxed{p(\mathbf{e} \mid c, \mathbf{x})}}$$



Posterior · Prior · Likelihood

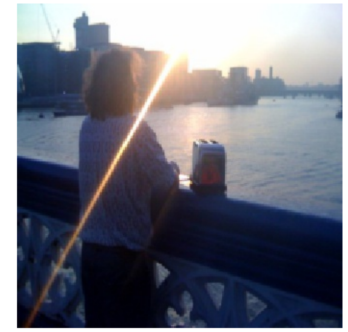
$$\boxed{P(c \mid \mathbf{e}, \mathbf{x})} \propto \boxed{P(c \mid \mathbf{x})}\,\boxed{p(\mathbf{e} \mid c, \mathbf{x})}$$

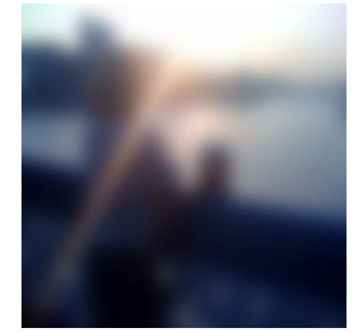
Which category do you think the robot will classify the image as?

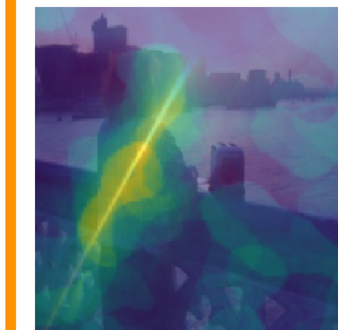


Toaster
Quill

Enclose the critical regions for classifying this image as Quill



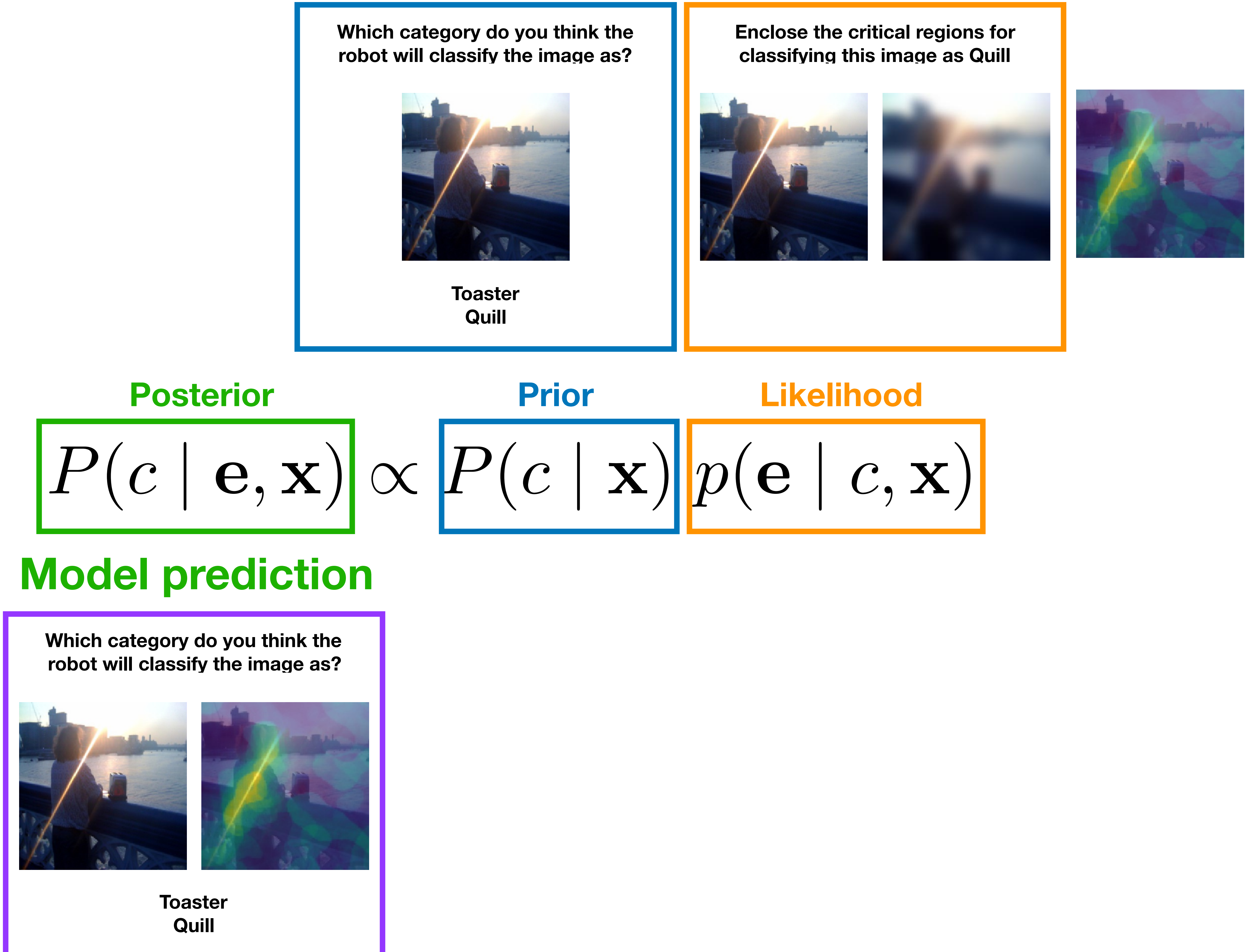




Posterior **Prior** **Likelihood**

$$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$$

Model prediction

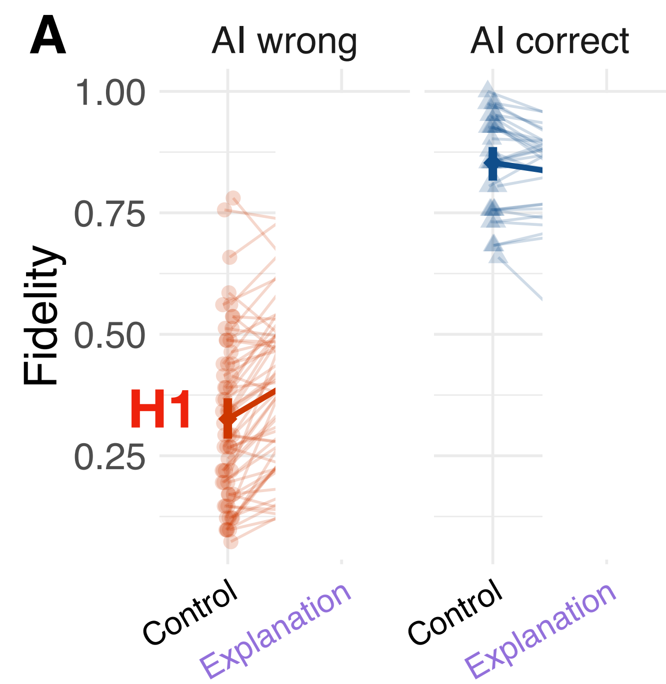

Which category do you think the robot will classify the image as?
Toaster
Quill
Enclose the critical regions for classifying this image as Quill
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Model prediction
Which category do you think the robot will classify the image as?
Toaster
Quill

Results

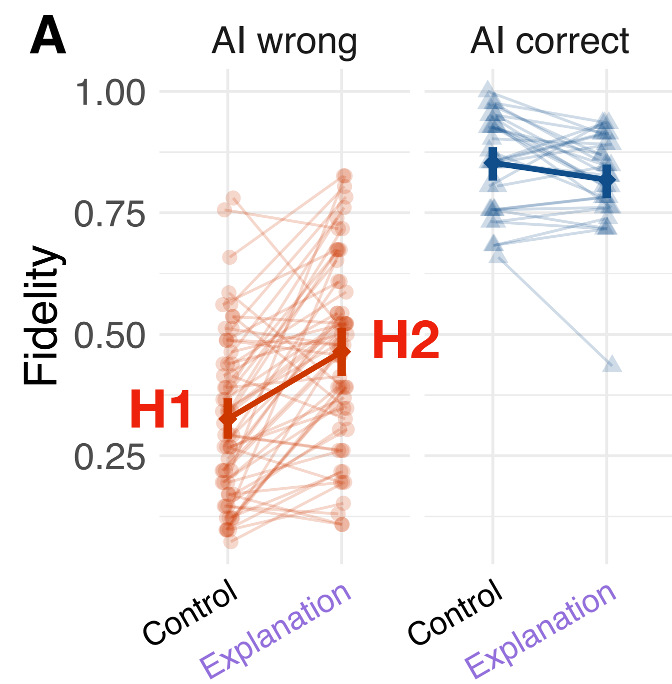
Fidelity:

probability that participants correctly predict the AI classification

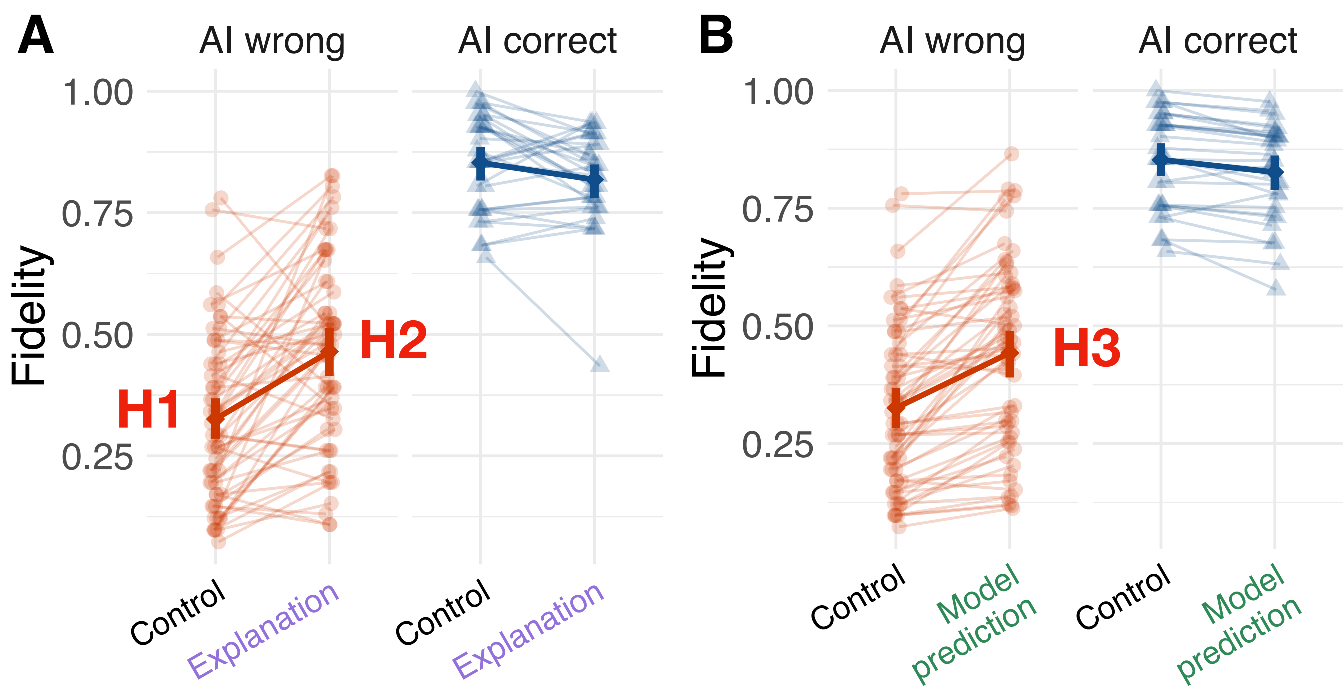
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.



1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).



1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.



LOO-CV MSE:

Leave-one-out cross validation: standard way to compare models with different parameterizations

Mean squared error: discrepancy between the participants response and model prediction

Does the likelihood matter?

Full model:

$$\overbrace{P(c \mid \mathbf{e}, \mathbf{x})}^{\text{Posterior}} \propto \overbrace{P(c \mid \mathbf{x})}^{\text{Prior}} \; \overbrace{p(\mathbf{e} \mid c, \mathbf{x})}^{\text{Likelihood}}$$

vs

Prior-only model:

$$\overbrace{P(c \mid \mathbf{e}, \mathbf{x})}^{\text{Posterior}} \propto \overbrace{P(c \mid \mathbf{x})}^{\text{Prior}}$$

Does the psychological space matter?

Psychological distance
based on **pixel-wise difference**

$$sim[\mathbf{e}\,,\,\mathbf{e}'] = |\mathbf{e} - \mathbf{e}'|_1$$

L-1 model

VS

Psychological distance
based on **feature overlap**

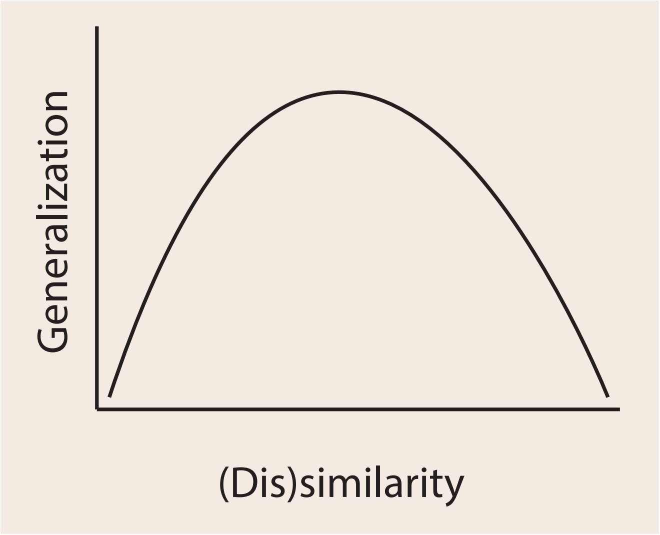
$$sim[\mathbf{e}\,,\,\mathbf{e}] = \frac{\langle \mathbf{e}\ ,\ \mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$$

Full model

Does the generalization function matter?

Non-monotonic generalization

$$p(\mathbf{e} \mid c, \mathbf{x}) = \frac{\Gamma(2\lambda)}{\Gamma(\lambda)\Gamma(\lambda)} sim[\mathbf{e}\,,\ \mathbf{e}']^{\lambda-1} \times (1 - sim[\mathbf{e}\,,\ \mathbf{e}'])^{\lambda-1}$$

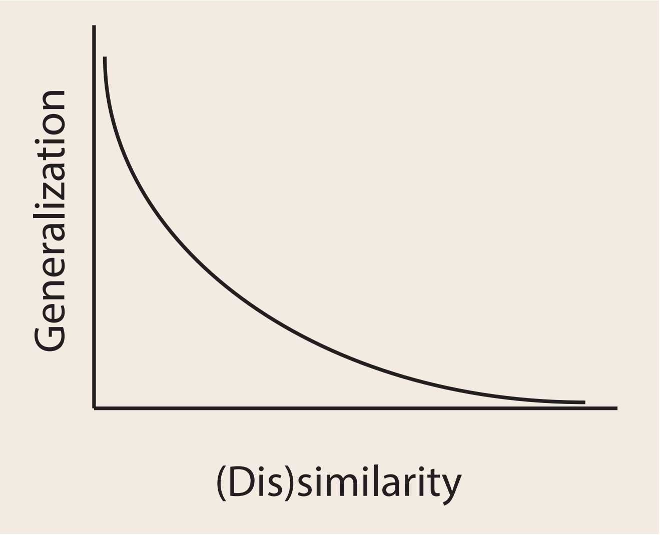


Beta model

VS

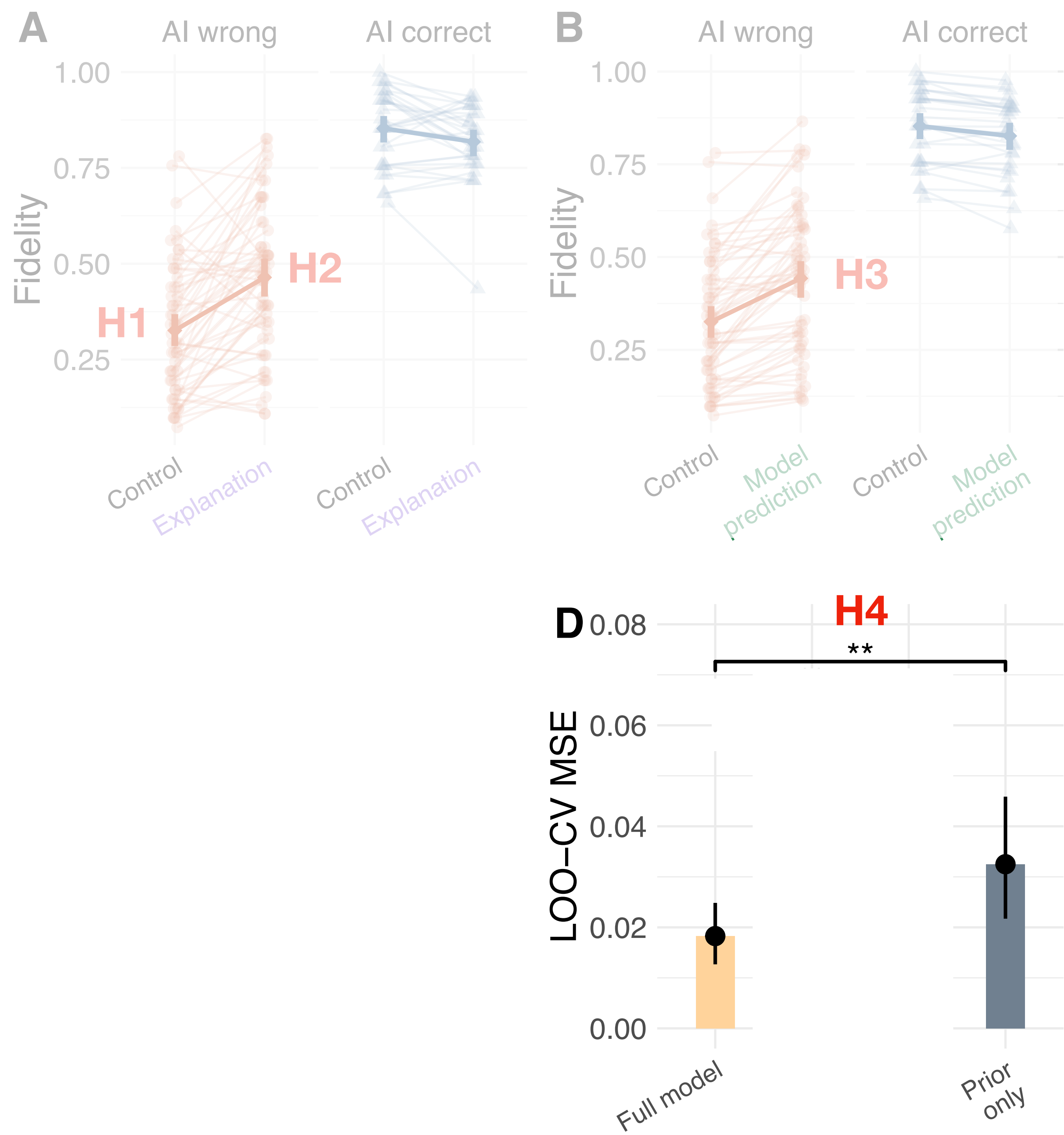
Monotonic generalization

$$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp[-\lambda\ (1 - sim[\mathbf{e}\,,\ \mathbf{e}'])]$$

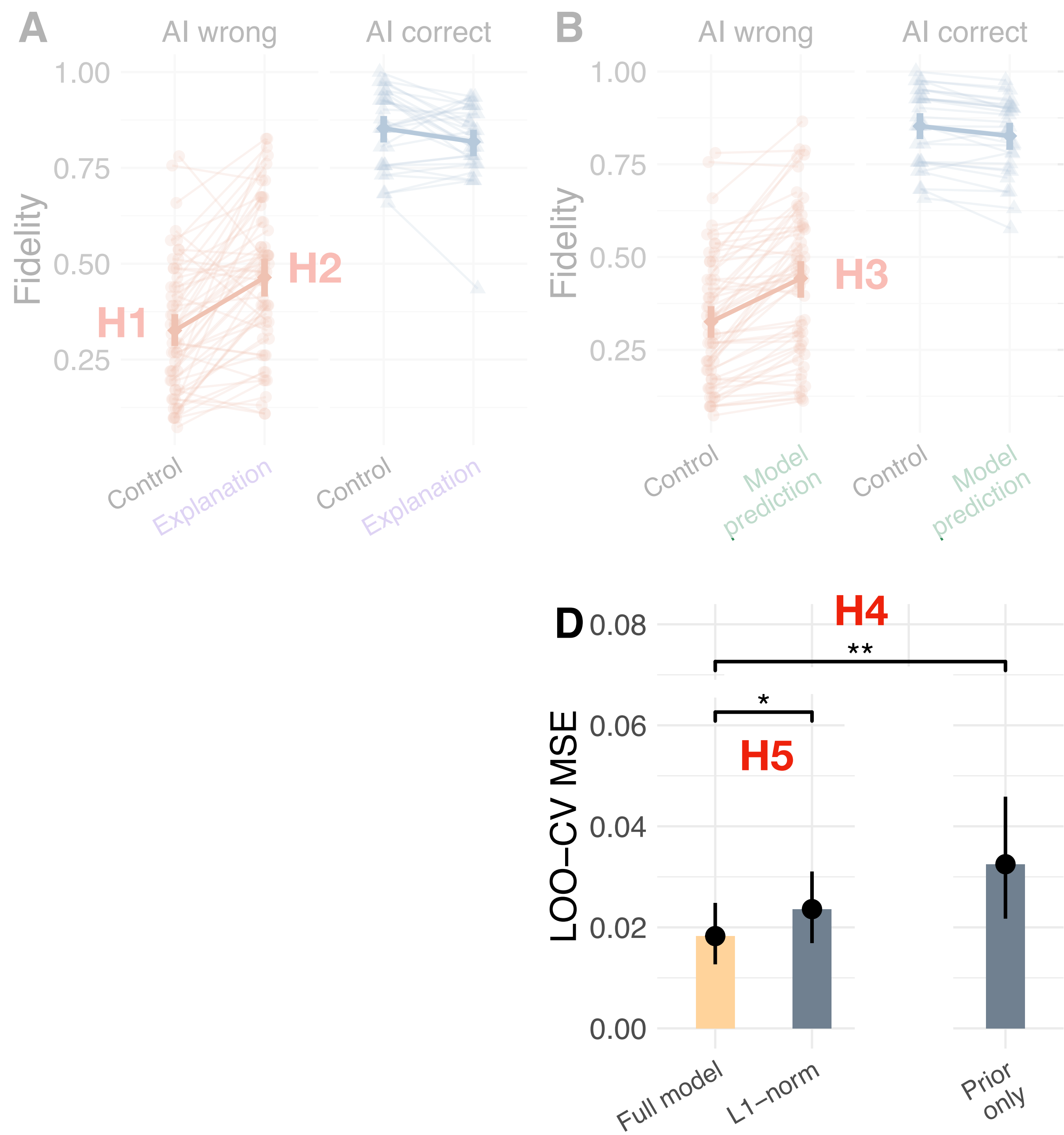


Full model

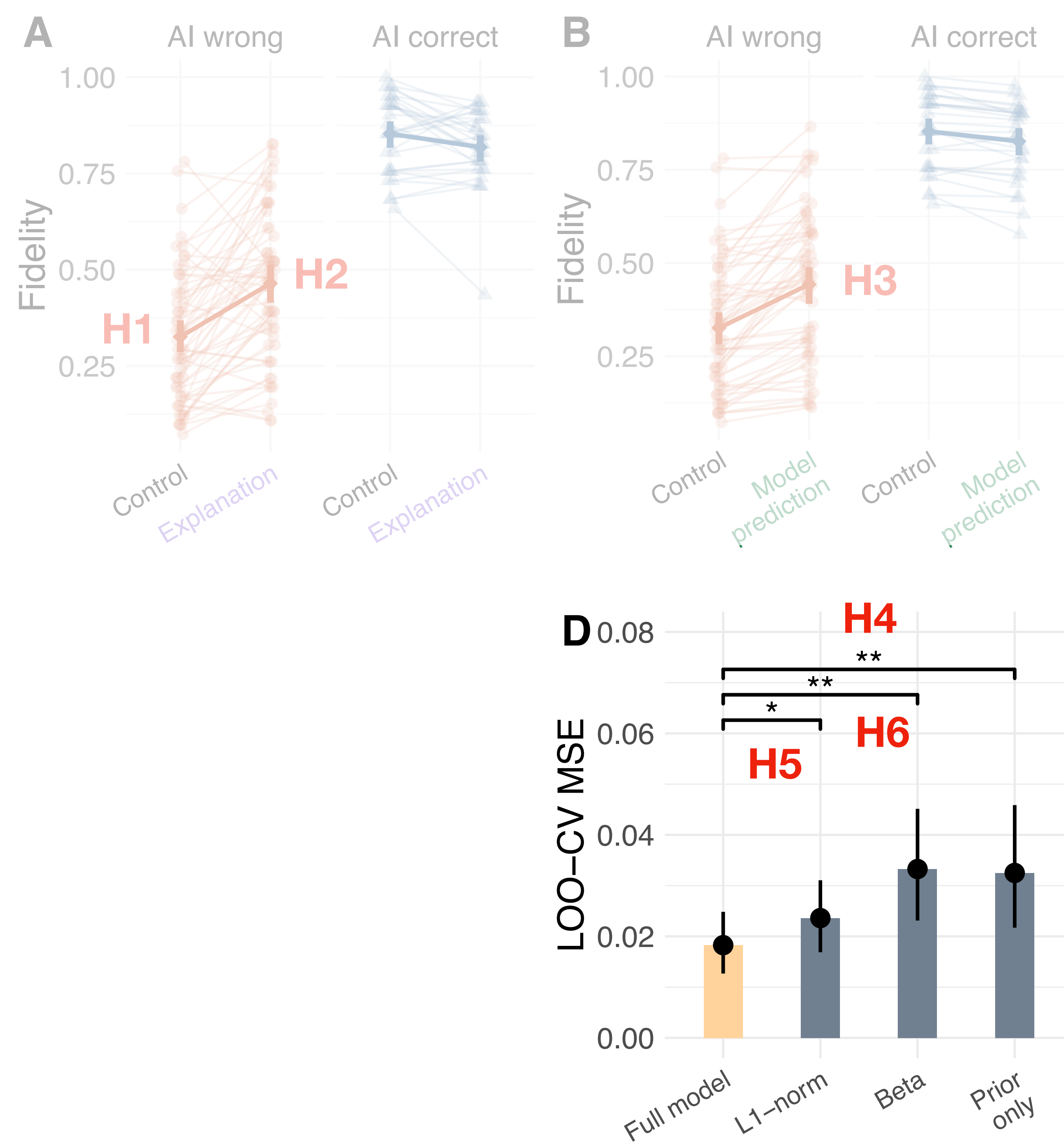
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.
4. The likelihood captures belief-updating from specific explanations, meaning that the full model is better than a prior-only model at predicting human behavior.



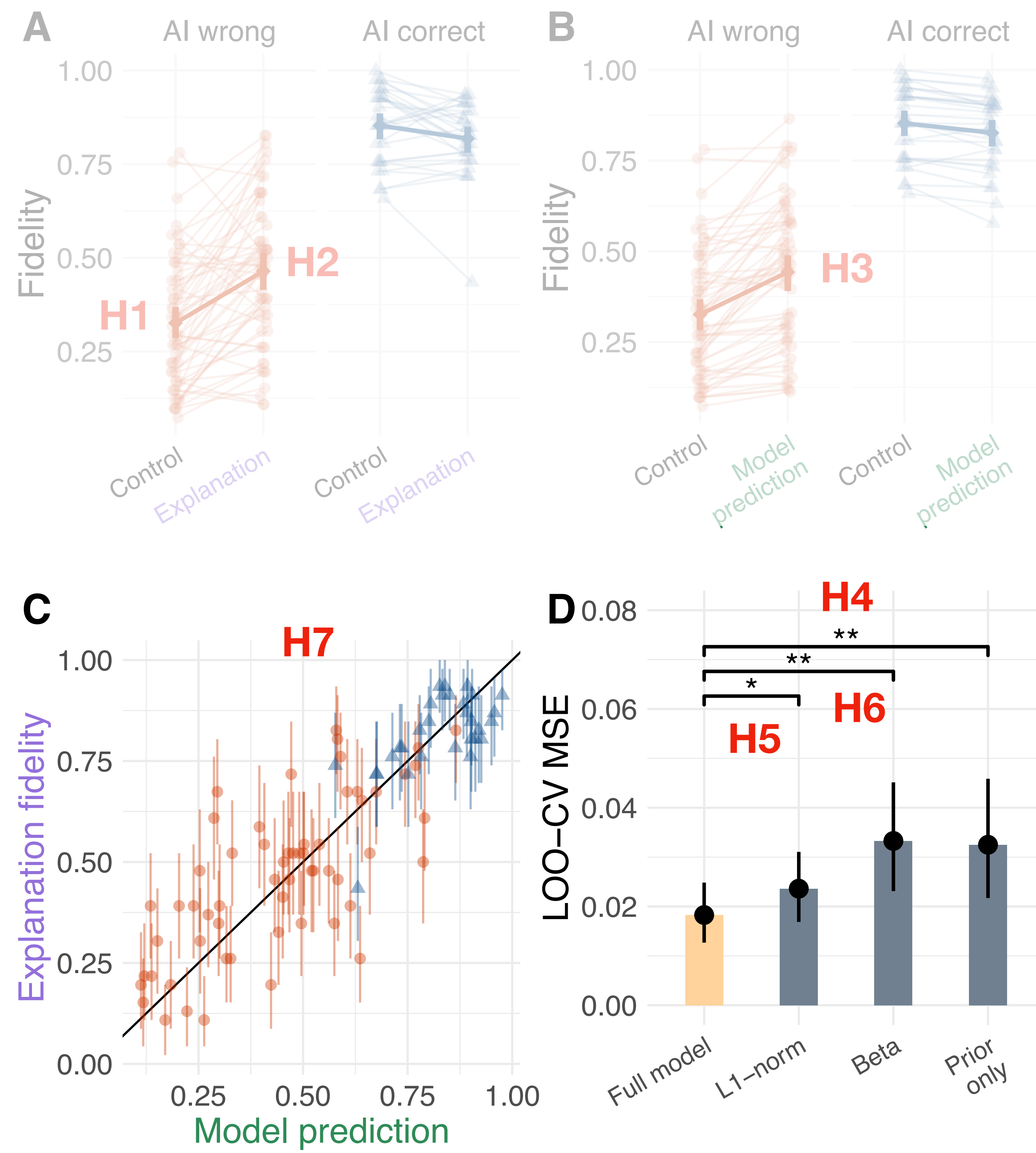
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.
4. The likelihood captures belief-updating from specific explanations, meaning that the full model is better than a prior-only model at predicting human behavior.
5. Comparison between explanations is done in a psychological space, implying that less-natural space (L1-norm) will be worse.



1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.
4. The likelihood captures belief-updating from specific explanations, meaning that the full model is better than a prior-only model at predicting human behavior.
5. Comparison between explanations is done in a psychological space, implying that less-natural space (L1-norm) will be worse.
6. Generalization follows Shepard's universal law and decays monotonically with increasing psychological distance, implying that distributions that violate this decay (Beta(λ,λ)) will be worse.



1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.
4. The likelihood captures belief-updating from specific explanations, meaning that the full model is better than a prior-only model at predicting human behavior.
5. Comparison between explanations is done in a psychological space, implying that less-natural space (L1-norm) will be worse.
6. Generalization follows Shepard's universal law and decays monotonically with increasing psychological distance, implying that distributions that violate this decay (Beta(λ,λ)) will be worse.
7. The theory can predict human response across a wide range of stimuli, classes, and explanations.



Contributions

- ★ Psychological theory of explainability
  - ✦ Humans project their own belief onto the AI
  - ✦ Effective explanations mitigate this belief projection
  - ✦ Humans interpret a received explanation by comparing it to self-generated explanations
    - ✤ The comparison occurs in a suitable psychological space
    - ✤ The comparison is turned to a response follows Shepard's universal law of generalization