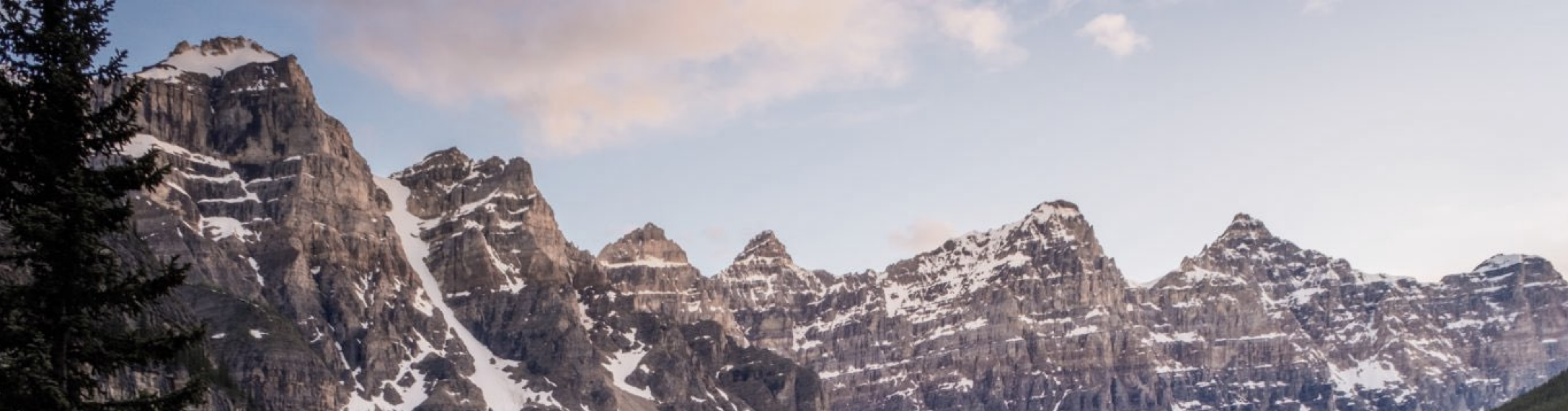
A Psychological Theory of Explainability

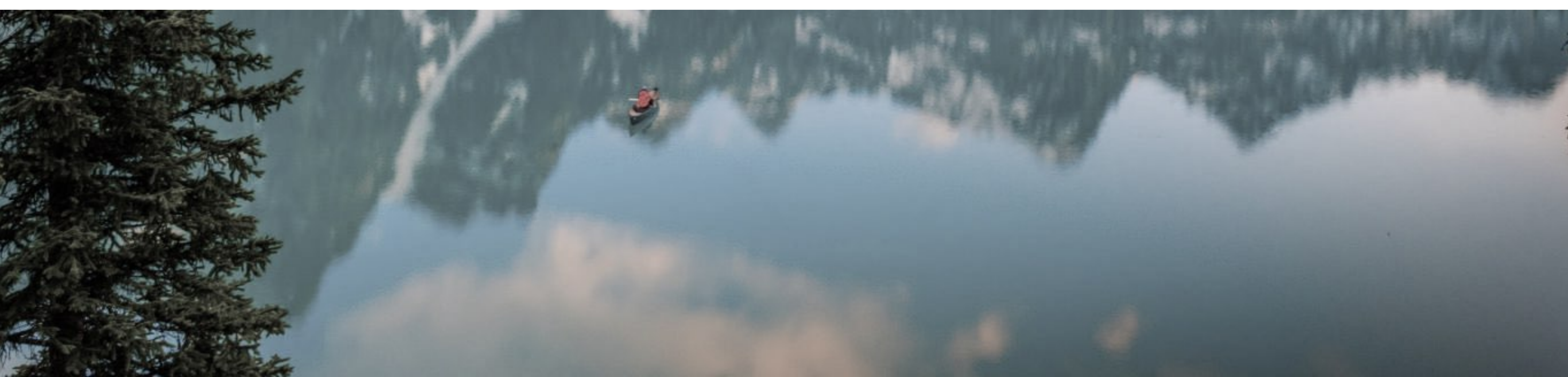
Scott Cheng-Hsin Yang*, Tomas Folke* & Patrick Shafto
*equal contribution

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

- ✔ Techniques to generate explanations
- ✔ Analysis of the techniques
- ✔ Validation of the techniques
- ✘ How humans interpret the explanations given

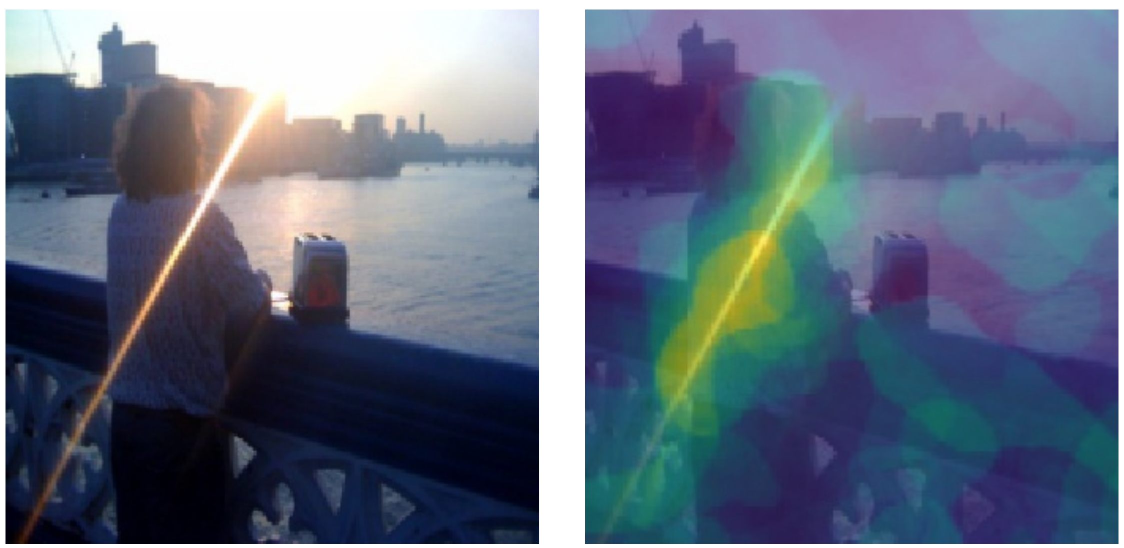


Humans **project their beliefs** onto the AI; thus, they interpret the explanation provided by **comparing it to the explanations that they themselves would give.**



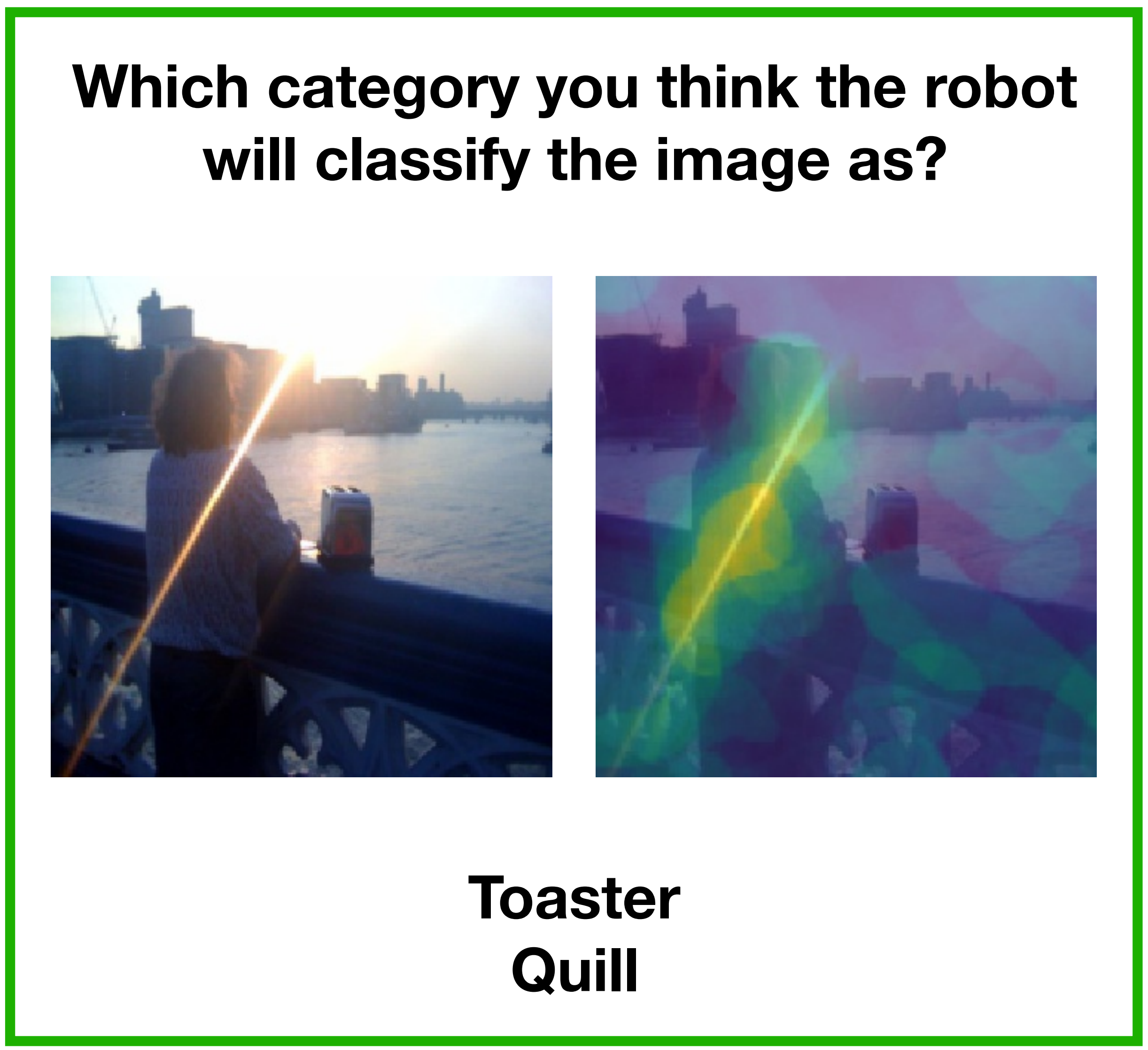
Example trial
(Explanation condition)

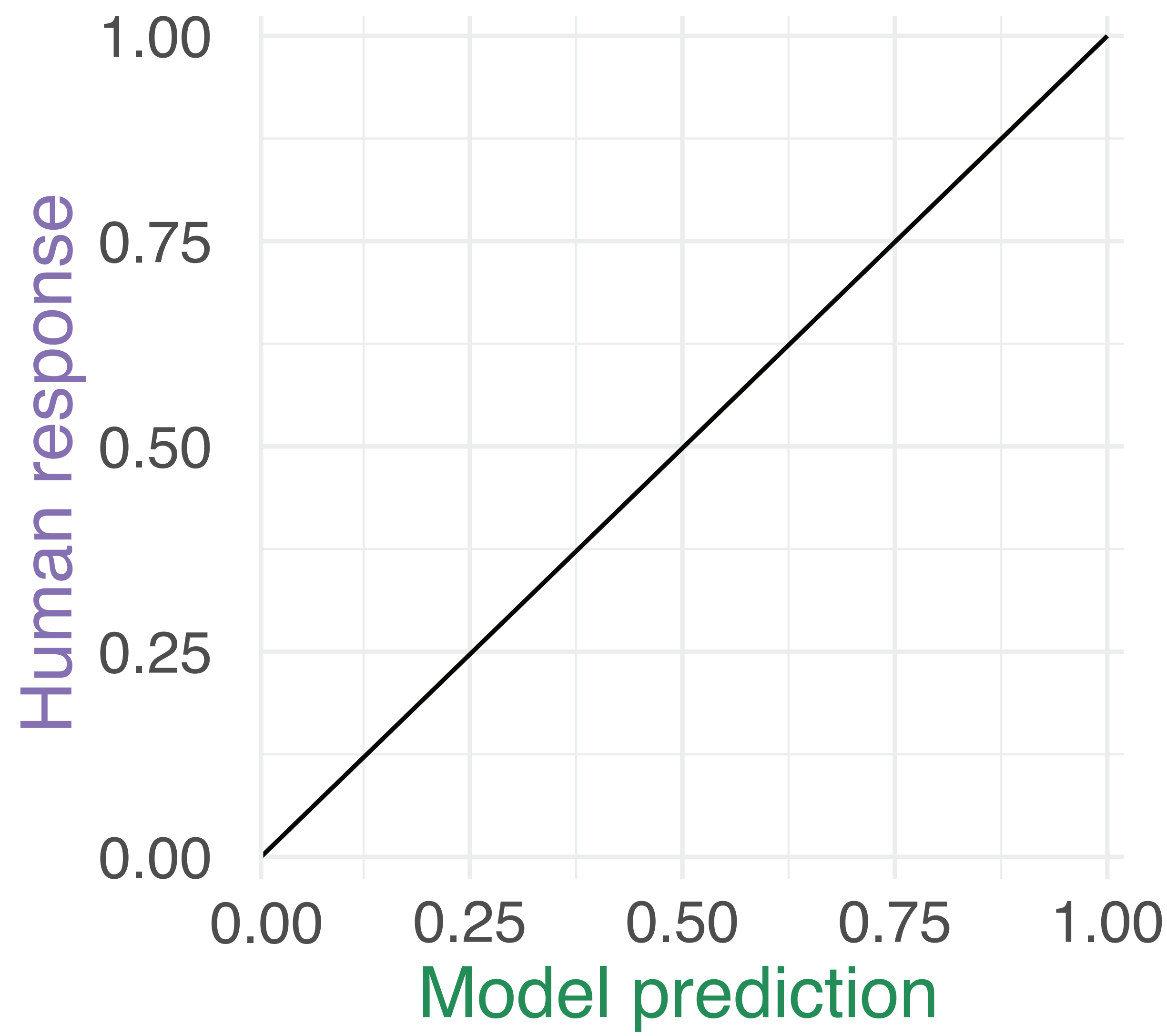
Which category you think the robot will classify the image as?



Toaster
Quill

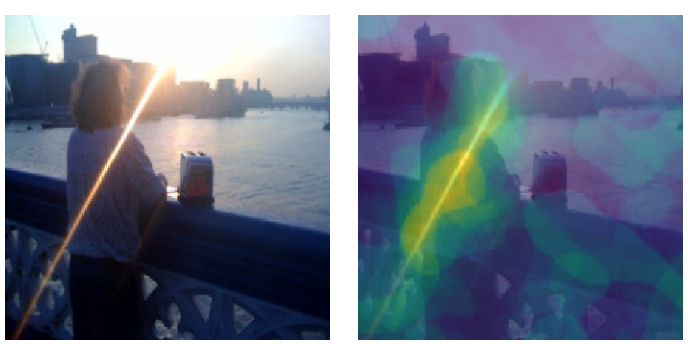
Example trial
(Explanation condition)



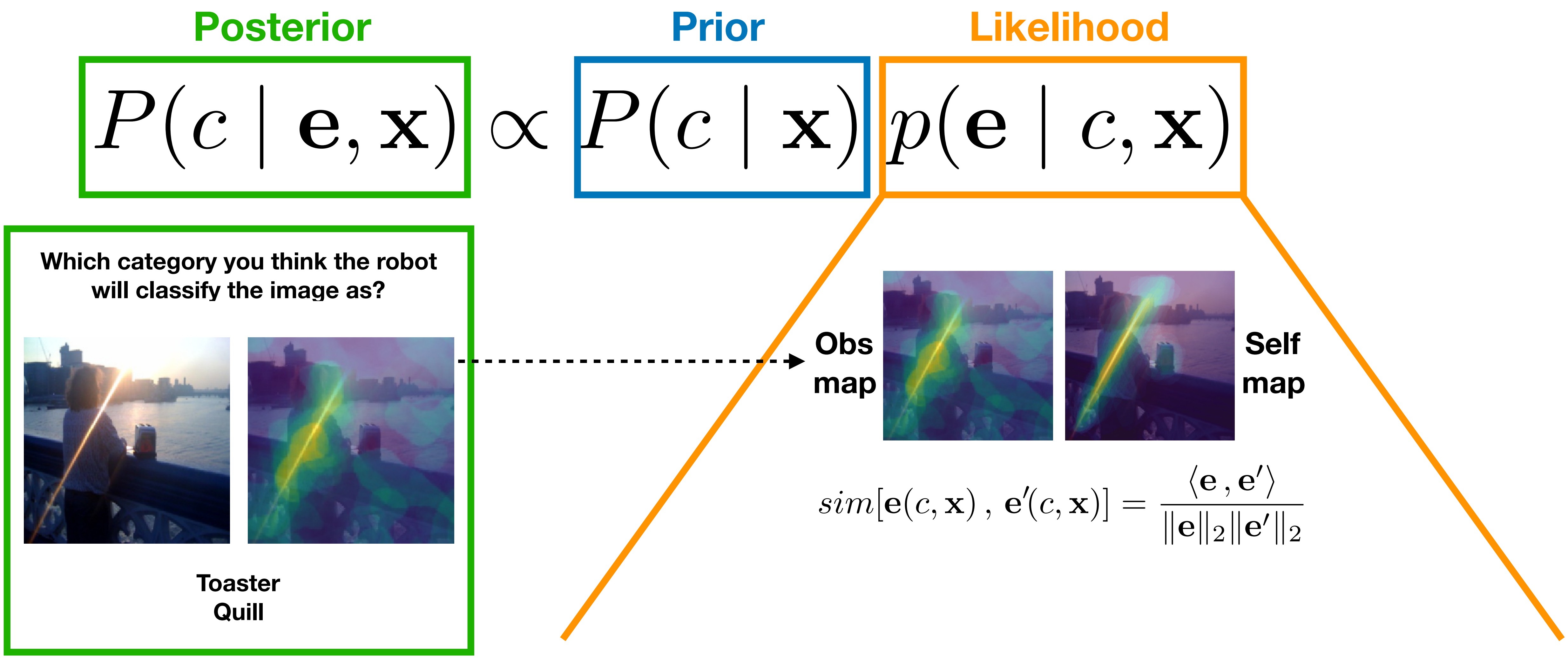


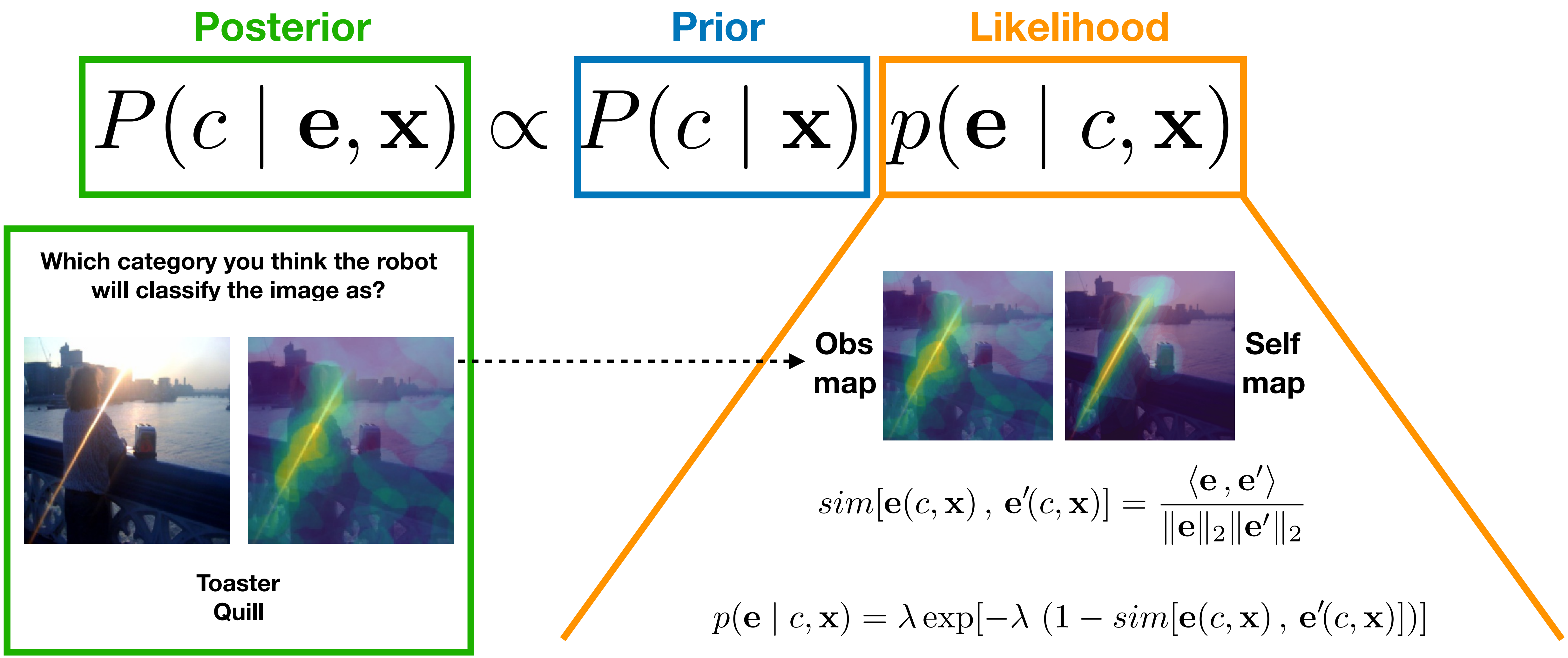
$$\overset{\textbf{Posterior}}{\boxed{P(c \mid \mathbf{e}, \mathbf{x})}} \propto \overset{\textbf{Prior}}{\boxed{P(c \mid \mathbf{x})}} \overset{\textbf{Likelihood}}{\boxed{p(\mathbf{e} \mid c, \mathbf{x})}}$$

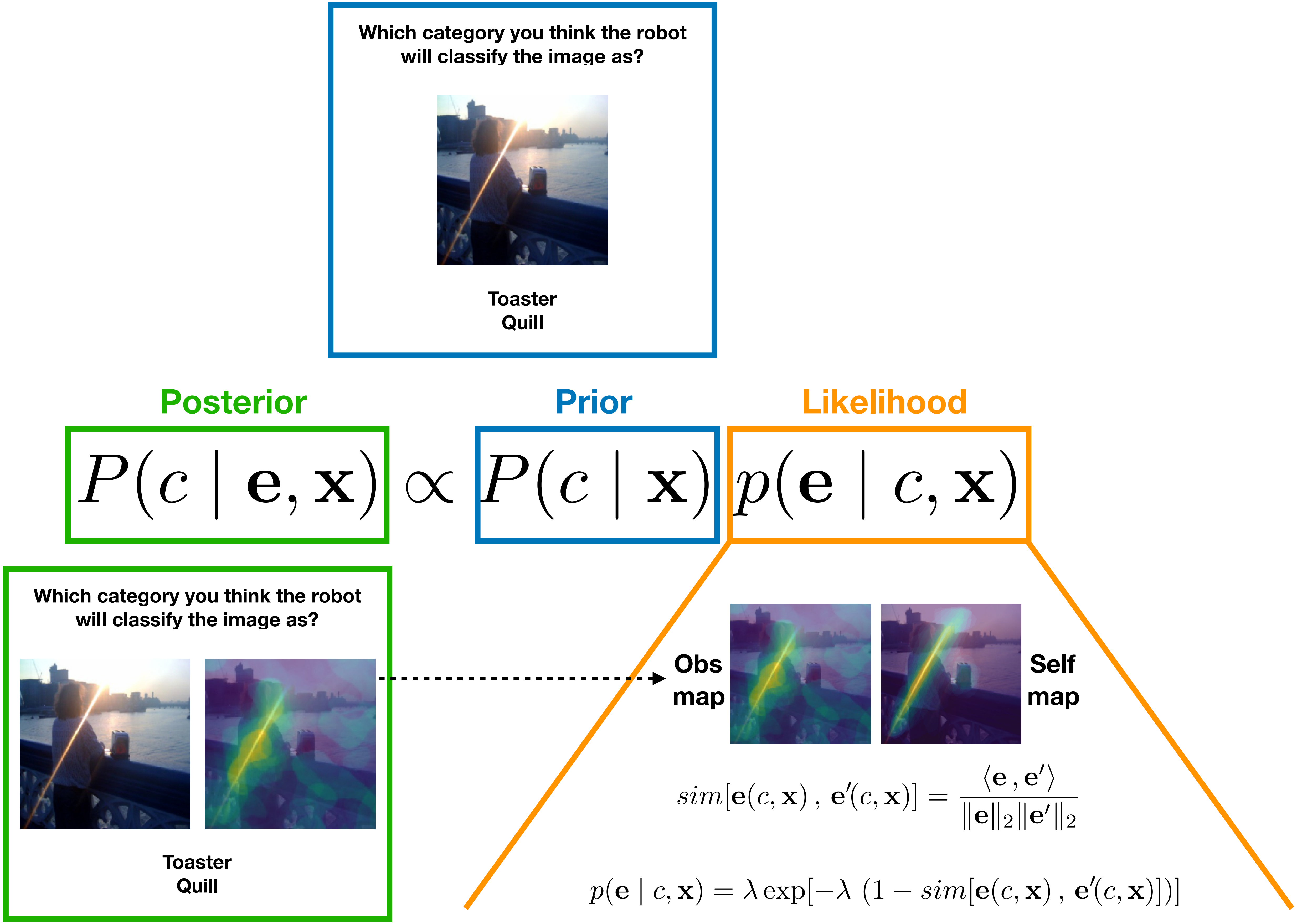
Which category you think the robot will classify the image as?

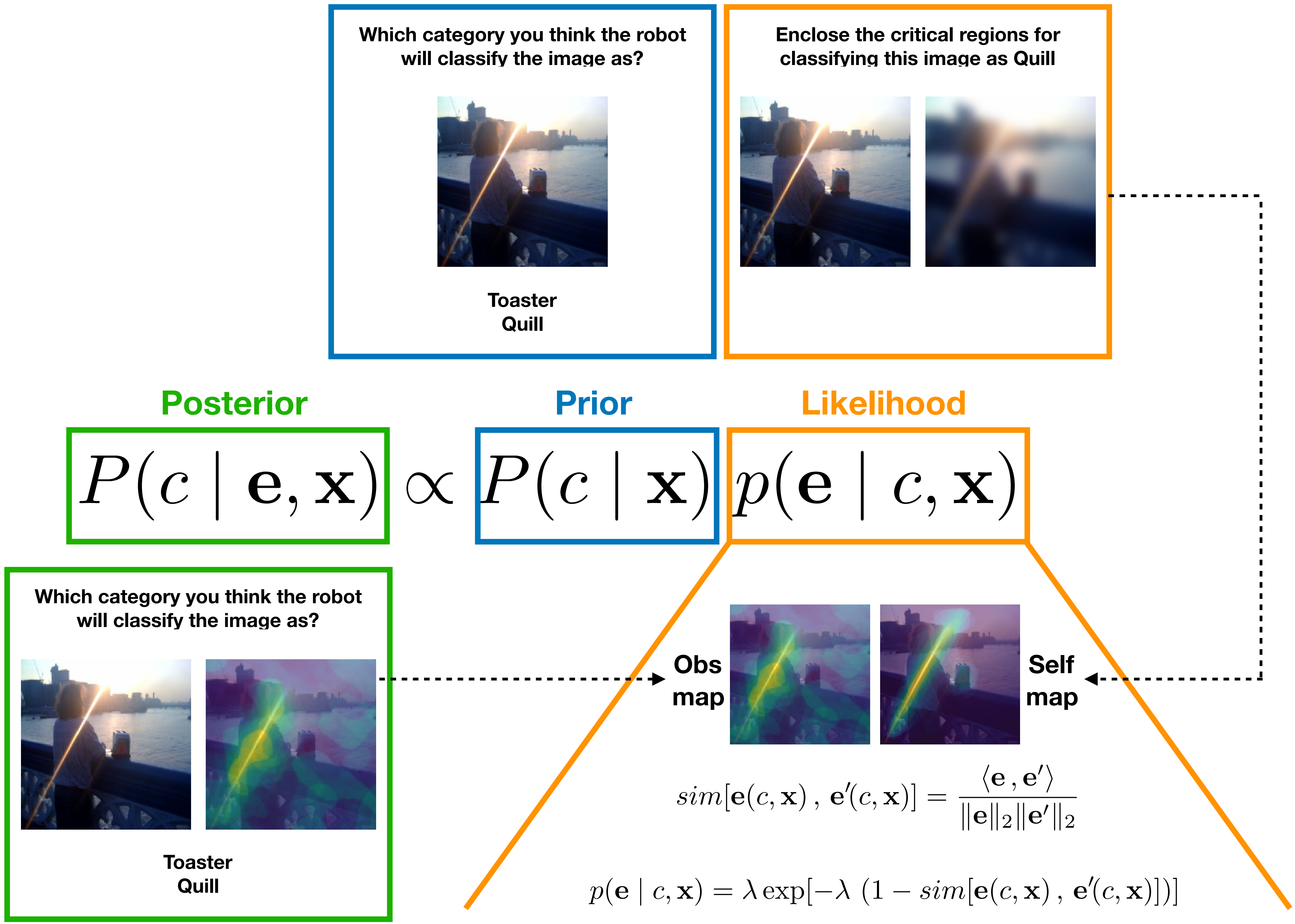


Toaster
Quill


Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Which category you think the robot will classify the image as?
Toaster
Quill
Obs map
Self map
$sim[\mathbf{e}(c,\mathbf{x}) \,,\, \mathbf{e}'(c,\mathbf{x})] = \frac{\langle \mathbf{e} \,,\, \mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$

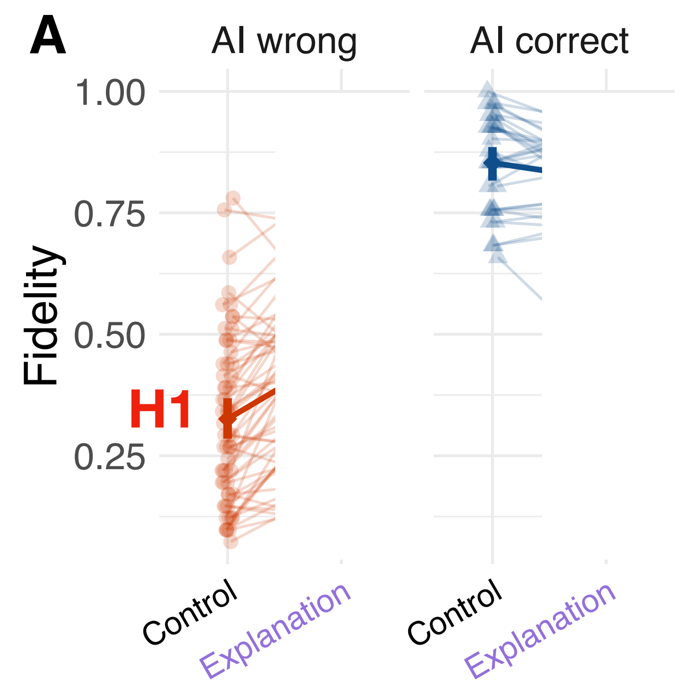

Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x}) \, p(\mathbf{e} \mid c, \mathbf{x})$
Which category you think the robot will classify the image as?
Toaster
Quill
Obs map
Self map
$sim[\mathbf{e}(c, \mathbf{x}) \, , \, \mathbf{e}'(c, \mathbf{x})] = \frac{\langle \mathbf{e} \, , \mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$
$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp[-\lambda \, (1 - sim[\mathbf{e}(c, \mathbf{x}) \, , \, \mathbf{e}'(c, \mathbf{x})])]$


Which category you think the robot will classify the image as?
Toaster
Quill
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Which category you think the robot will classify the image as?
Toaster
Quill
Obs map
Self map
$sim[\mathbf{e}(c,\mathbf{x})\,,\, \mathbf{e}'(c,\mathbf{x})] = \frac{\langle \mathbf{e}\,,\mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$
$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp[-\lambda\ (1 - sim[\mathbf{e}(c,\mathbf{x})\,,\, \mathbf{e}'(c,\mathbf{x})])]$

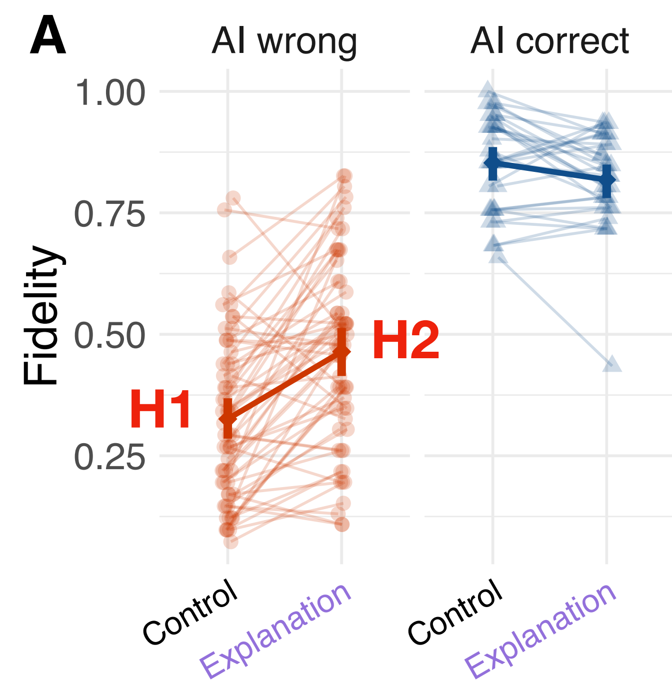


Which category you think the robot will classify the image as?
Toaster
Quill
Enclose the critical regions for classifying this image as Quill
Posterior
Prior
Likelihood
$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\, p(\mathbf{e} \mid c, \mathbf{x})$
Which category you think the robot will classify the image as?
Toaster
Quill
Obs map
Self map
$sim[\mathbf{e}(c,\mathbf{x})\,,\,\mathbf{e}'(c,\mathbf{x})] = \frac{\langle \mathbf{e}\,,\mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$
$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp[-\lambda\ (1 - sim[\mathbf{e}(c,\mathbf{x})\,,\,\mathbf{e}'(c,\mathbf{x})])]$

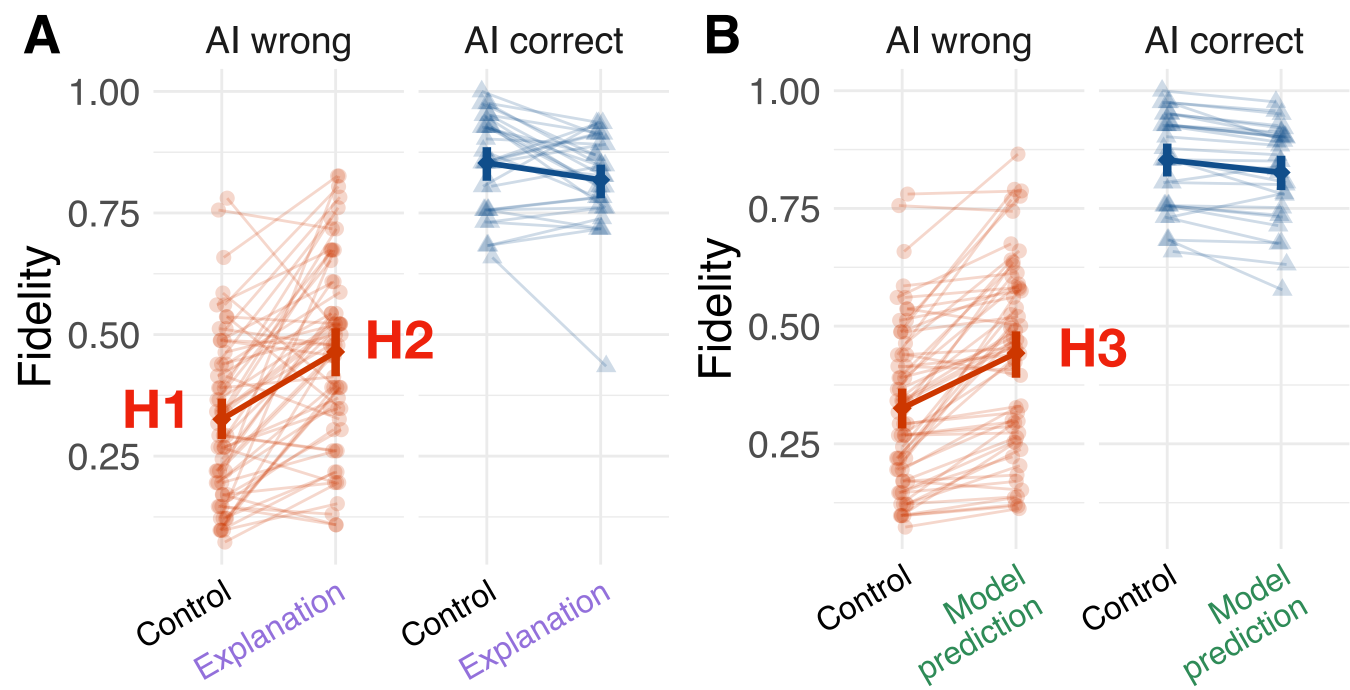
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.



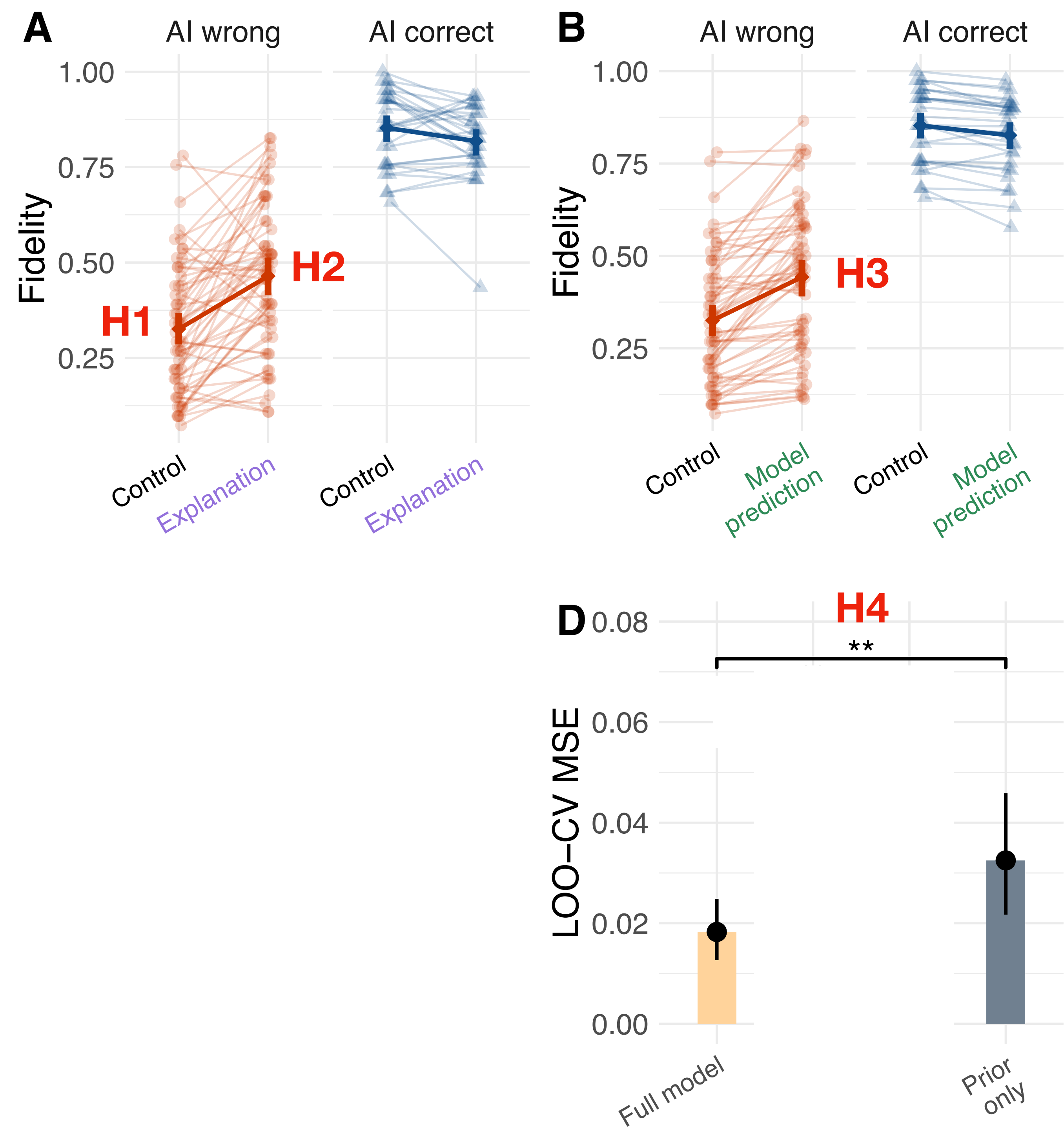
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).



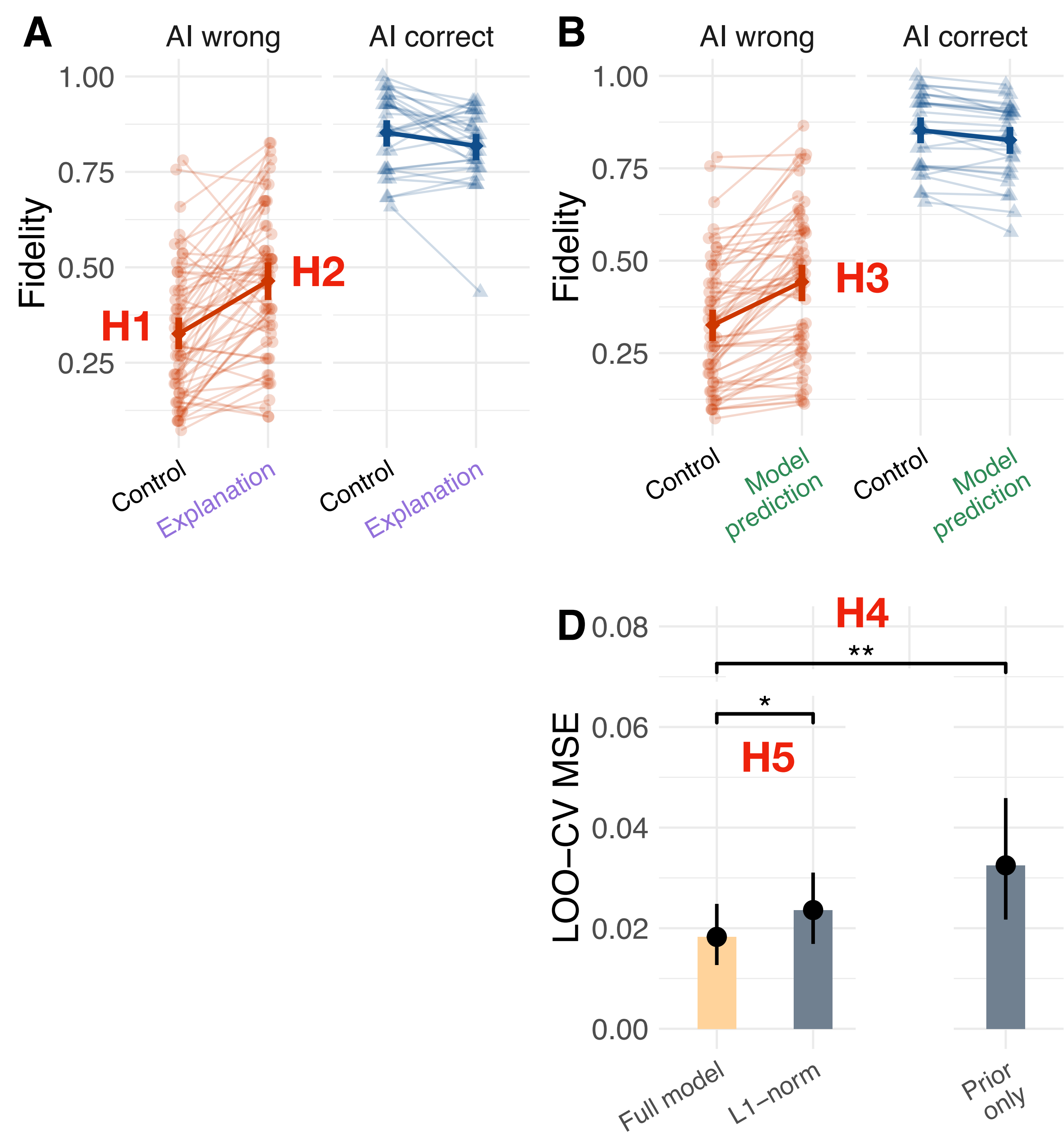
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.



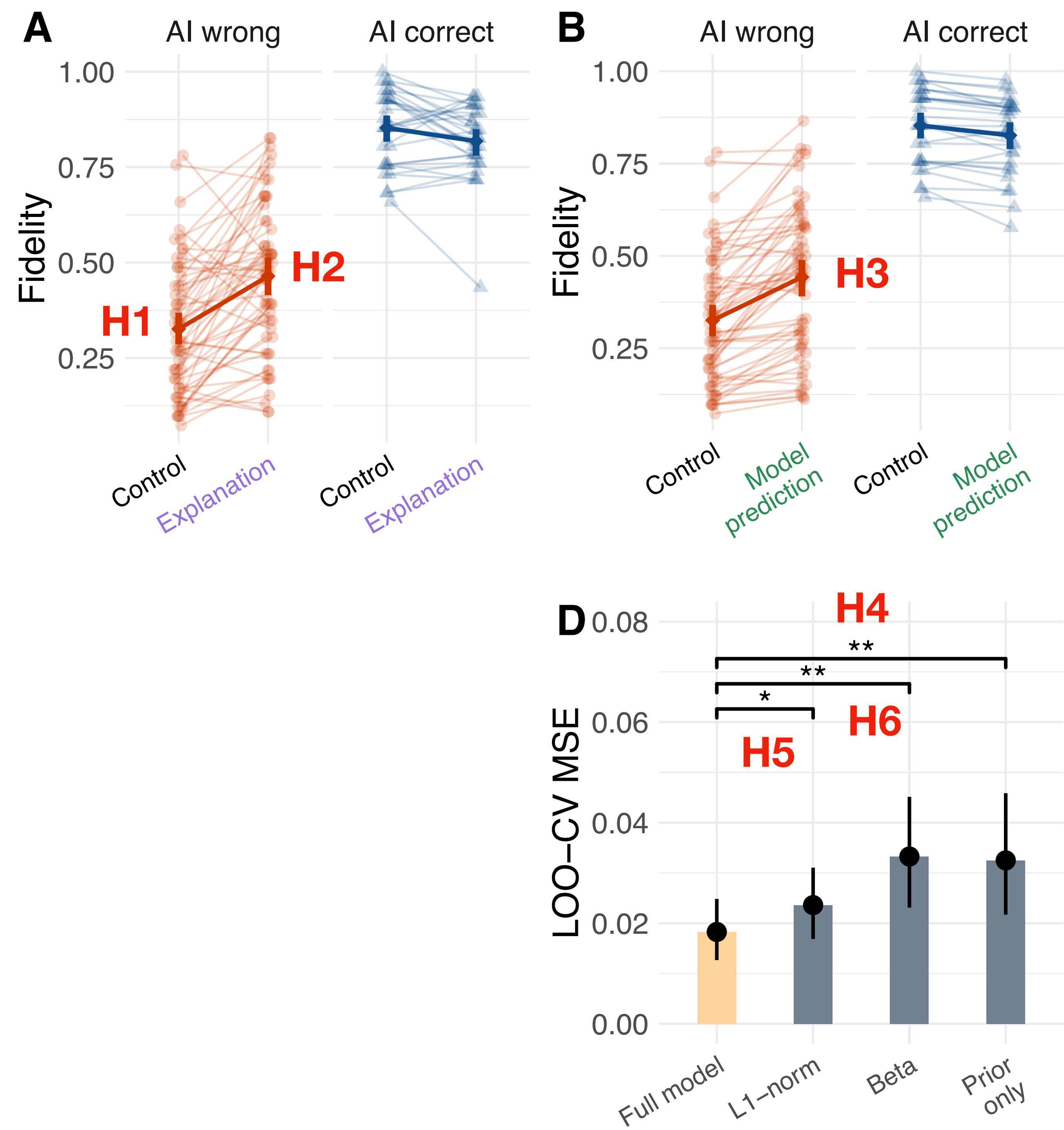
1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong.
2. Good explanations increase fidelity, especially when the original fidelity is low (when AI is wrong).
3. Model prediction recovers H2.
4. The 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 predicts human response well across a wide range of stimuli, classes, and explanations.