# Explanation Alignment

## Quantifying the Correctness of Model Reasoning At Scale


**Hyemin Bang**
hbang@csail.mit.edu
@hhybang2

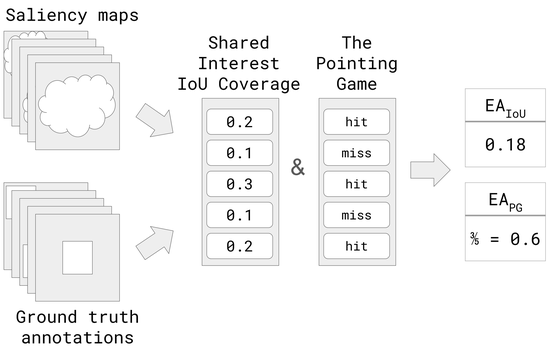
**Angie Boggust**
aboggust@csail.mit.edu
@angie_boggust

**Arvind Satyanarayan**
arvindsatya@mit.edu
@arvindsatya1


**Explanation Alignment** measures the human alignment of model decisions.

We aggregate saliency-based metrics, Shared Interest and The Pointing Game, to quantify how often models make decisions for the right reasons.

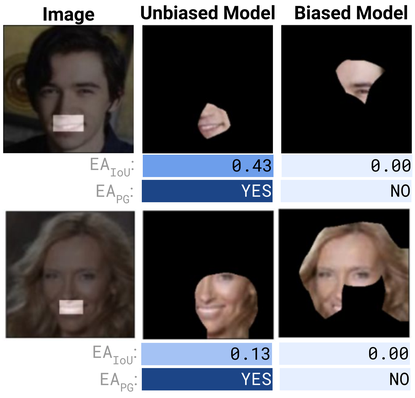
- $EA_{IoU}$ measures the overlap between human and model explanations.
- $EA_{PG}$ indicates how often the model's most important feature aligns with human expectations.

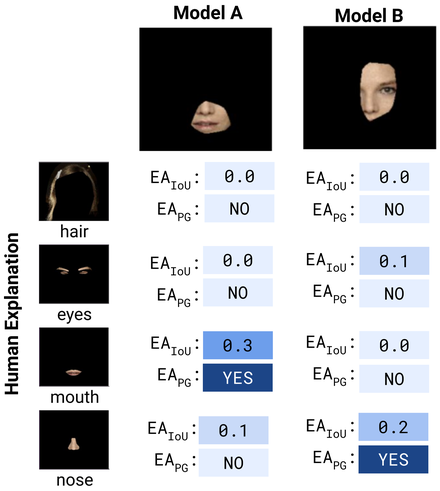



**Revealing Model Bias in Face Classification Models** without previous knowledge of possible bias or manual test procedures.

In the CelebA smile prediction task, the `biased` model shows low alignment with the mouth, indicating reliance on other features, while the `unbiased` model aligns correctly with the mouth.

| Model Type | Accuracy on Data Splits | | | | Explanation Alignment | |
|---|---|---|---|---|---|---|
| | 👩 🙂 | 👨 😐 | 👨 🙂 | 👩 😐 | $EA_{IoU}$ | $EA_{PG}$ |
| unbiased | 91% | 91% | 92% | 95% | 0.18 | 0.26 |
| biased | 99% | 99% | 47% | 15% | 0.0 | 0.0 |




**Exposing Behavioral Differences in Highly Accurate Models** to make more informed decisions between models.

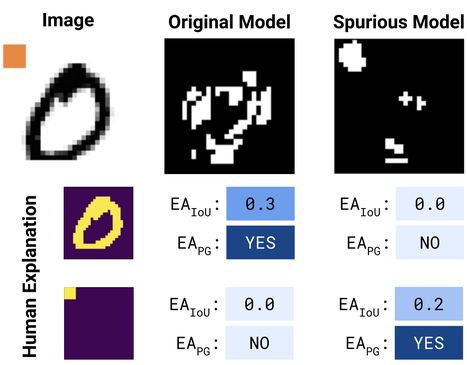
On CelebA smile prediction tasks, while model A heavily relies on features of the mouth, model B almost rarely relies on the mouth but more on other facial features.
- The choice of model can depend on not only its correctness but also on its alignment with expected reasoning.

| Model | Accuracy | Explanation Alignment ($EA_{IoU}$ / $EA_{PG}$) | | | |
|---|---|---|---|---|---|
| | | Hair | Eye | Mouth | Nose |
| A | 93% | 0.0 / 0.0 | 0.0 / 0.0 | **0.3 / 0.8** | 0.1 / 0.0 |
| B | 92% | 0.0 / 0.0 | 0.1 / 0.0 | 0.0 / 0.0 | **0.2 / 0.5** |

**Uncovering Spurious Correlations in a Controlled Setting** by augmenting MNIST dataset with color box with spurious correlation between the color and the digit.

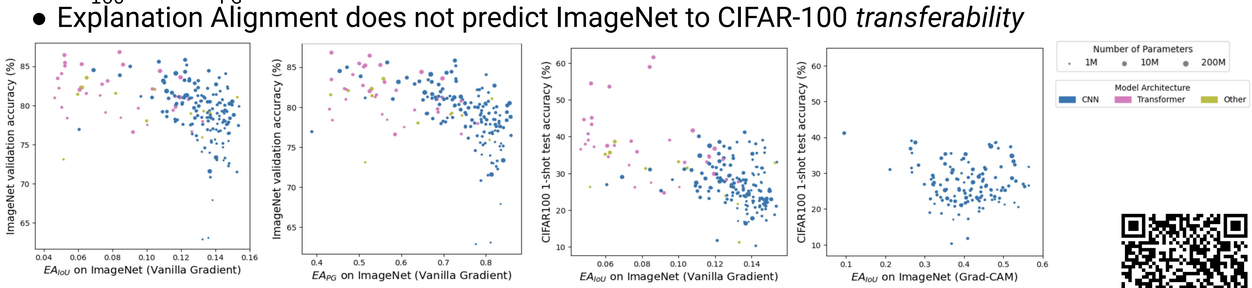
The `spurious` model learns the correlation, shown with its dependence on the color box, unlike `not-spurious` model relying on the digit.

| Model Type | Test Set Accuracy | | Digit | | Color Box | |
|---|---|---|---|---|---|---|
| | `not-spurious` | `spurious` | $EA_{IoU}$ | $EA_{PG}$ | $EA_{IoU}$ | $EA_{PG}$ |
| `not-spurious` | 98% | 98% | 0.3 | 0.7 | 0.0 | 0.0 |
| `spurious` | 46% | 100% | 0.1 | 0.0 | 0.2 | 0.9 |




**Characterizing** explanation alignment of 195 models across saliency methods, explanation alignment metrics, and tasks.

- Explanation Alignment differs based on *model architecture*
- Explanation Alignment is sensitive to the underlying *saliency method*
- $EA_{IoU}$ and $EA_{PG}$ are *interchangeable* for relative model comparisons
- Explanation Alignment does not predict ImageNet to CIFAR-100 *transferability*





**Explanation Alignment: Quantifying the Correctness of Model Reasoning At Scale**
Hyemin Bang, Angie Boggust, and Arvind Satyanarayan
*eXCV Workshop at the European Conference on Computer Vision (ECCV), Milan, Italy*