**What do we want to ask humans?**

* General quality: do you think the explanation is good?
* Understandability: do you think the explanation helps you to understand the AI system decision?
* Trustworthiness: how well do you trust the AI system based on the explanation?
* Transparency: do you think the explanation transparently describes why and how the AI system made its decision?
* Binary preference: which explanation do you find (better, more trustworthy, more understandable, more transparent etc.) (show 2 on the screen in a randomized order and design)?

Make a list of "evaluation" angles: general goodness, understandability, trustworthiness... first study phase: show a single explanation at a time, ask questions

This will lead to 1500 screens

second study phase: show paired explanations (make ordering block design)

Five different explanations -> number of combinations

1) gpt(text) compare to shap,

2) gpt(text) compare to gpt(text+shap)...

--- all combinations and in both orders

Estimate the number of samples needed (how many cases with all combinations of explanations we show)

Comparisons:

gpt(text) vs. shap

gpt(text) vs. gpt(text+shap)

gpt(text) vs. gpt(text+classfier)

gtp(text) vs. gpt(text+classifier+shap)

shap vs. gpt(text)

...