Lecture 3 - finishing up from last time

The challenges with evaluating models

Ranjay Krishna | ranjay@cs.washington.edu

Challenges with evaluating models

#1: The replication crisis

#2: Labeling errors

#3: Generalization errors

#4: A static test dataset

#5: Distribution shifts

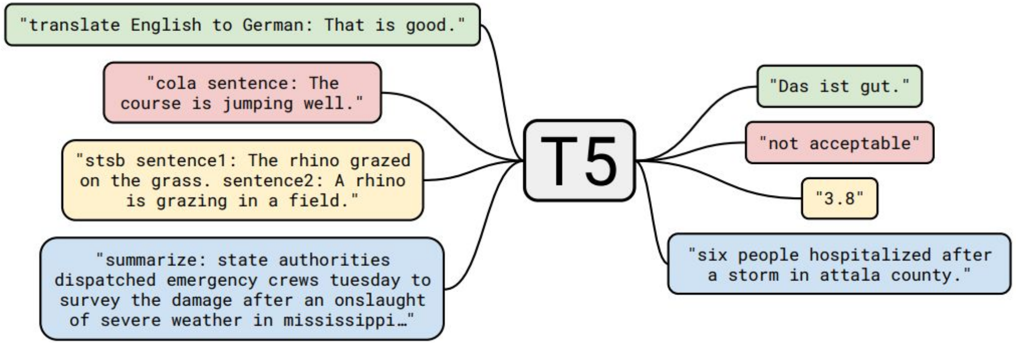
#6: Marginalization: Filtering

T5 trained on Colossal Clean Crawled Corpus

400 words from the [List of filtered words]

- E.g. swastika, white power - implications?
- E.g. twink - implications?

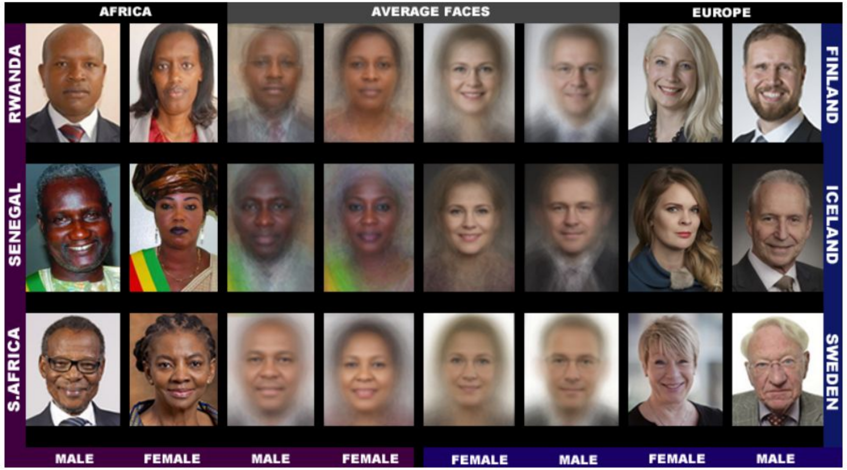
Dataset



Raffel et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. IJML 2020
Dodge et al. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. ArXiv 2021

#7: Bias in data source

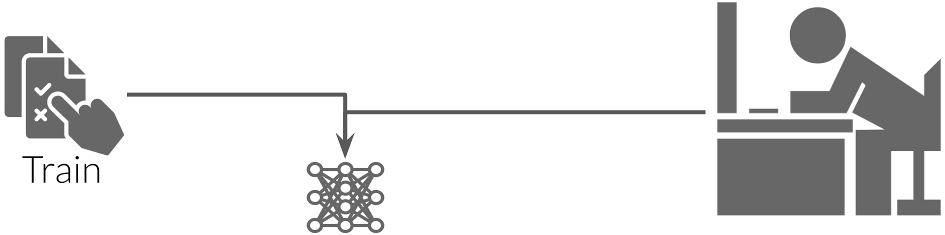
- Then: What was not curated caused bias
- Today: More media coverage = more training data instances



Data source

Buolamwini et al. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAccT 2018
Bender et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT 2021

#8: Environmental and financial costs



Energy for a flight from NY to SF:

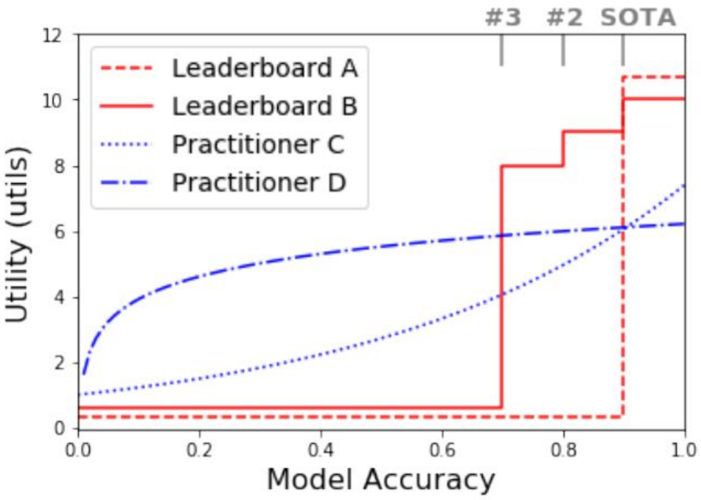
Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Strubell et al. Energy and Policy Considerations for Deep Learning in NLP. ACL 2019

#9: Leaderboard with one metric is not enough

Utility of a new AI model:

- is NOT smooth w.r.t. Accuracy for a leaderboard
- Any improvement along any dimension is good for a practitioner



Ethayarajh et al. Utility is in the Eye of the User: A Critique of NLP Leaderboards. EMNLP 2020

#10: Open ended tasks: Generative models are very hard to evaluate

Research question:

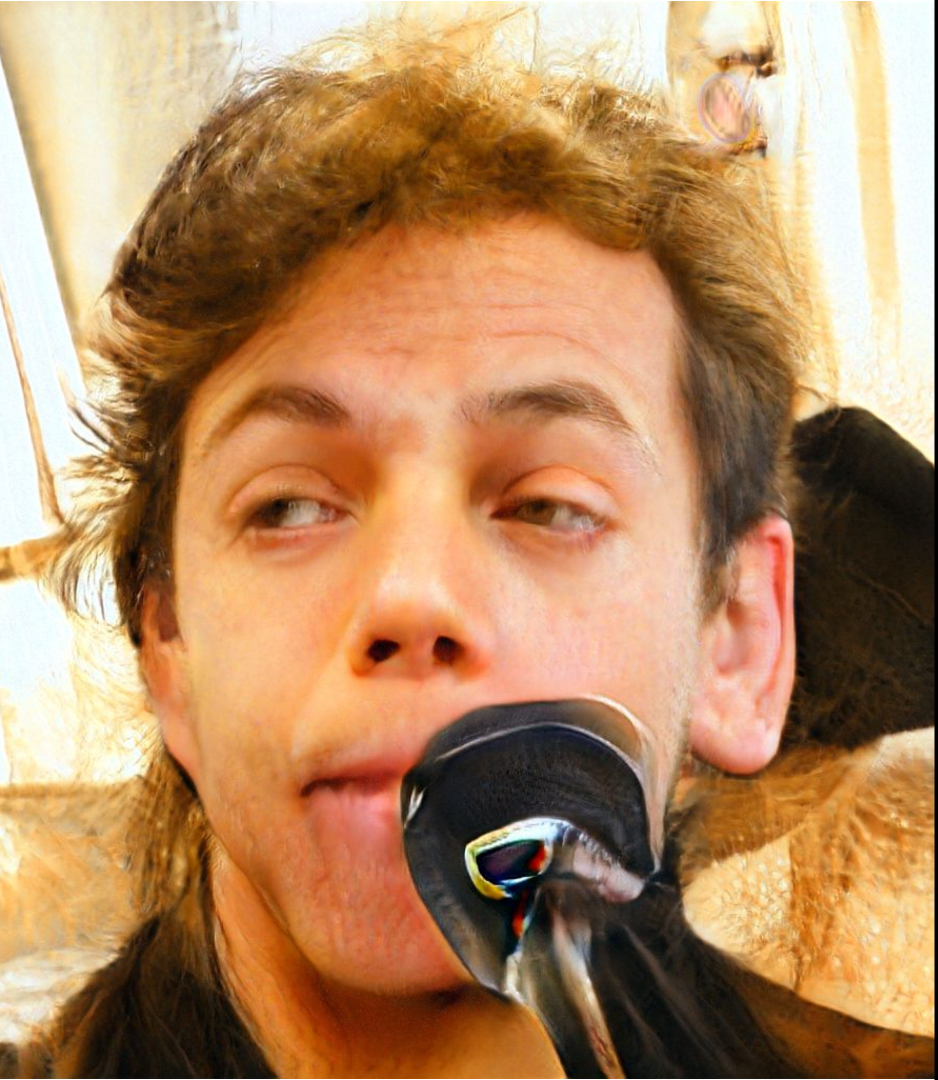
How do you evaluate the output of an image generation model?

Zhou et al. HYPE: A Benchmark For Human eYe Perceptual Evaluation of Generative Models. NeurIPS 2019












It used to be easy to measure progress

Ian Goodfellow @goodfellow_ian

Goodfellow, I. J., et al. "Generative Adversarial Networks." (2014).
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." (2015).
Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." (2016).
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." (2017).
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." (2019).

It’s much harder now

Ian Goodfellow @goodfellow_ian

Goodfellow, I. J., et al. "Generative Adversarial Networks." (2014).
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." (2015).
Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." (2016).
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." (2017).
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." (2019).

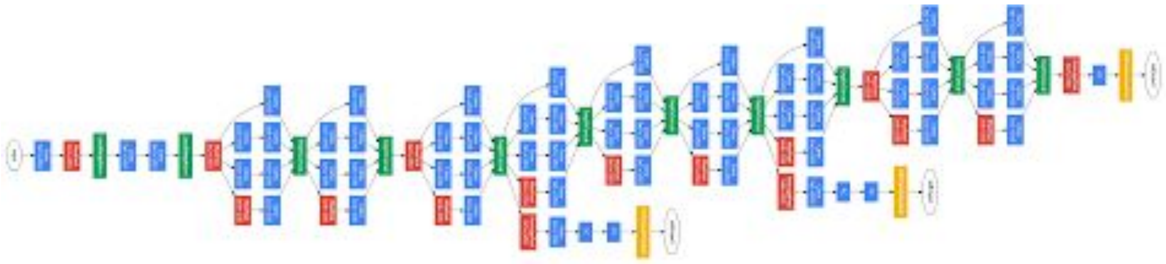
We don’t even have corresponding pairs

Ian Goodfellow @goodfellow_ian

Goodfellow, I. J., et al. "Generative Adversarial Networks." (2014).
Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." (2015).
Liu, Ming-Yu, and Oncel Tuzel. "Coupled generative adversarial networks." (2016).
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." (2017).
Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." (2019).

How are models evaluated today?

Inception score, FID.



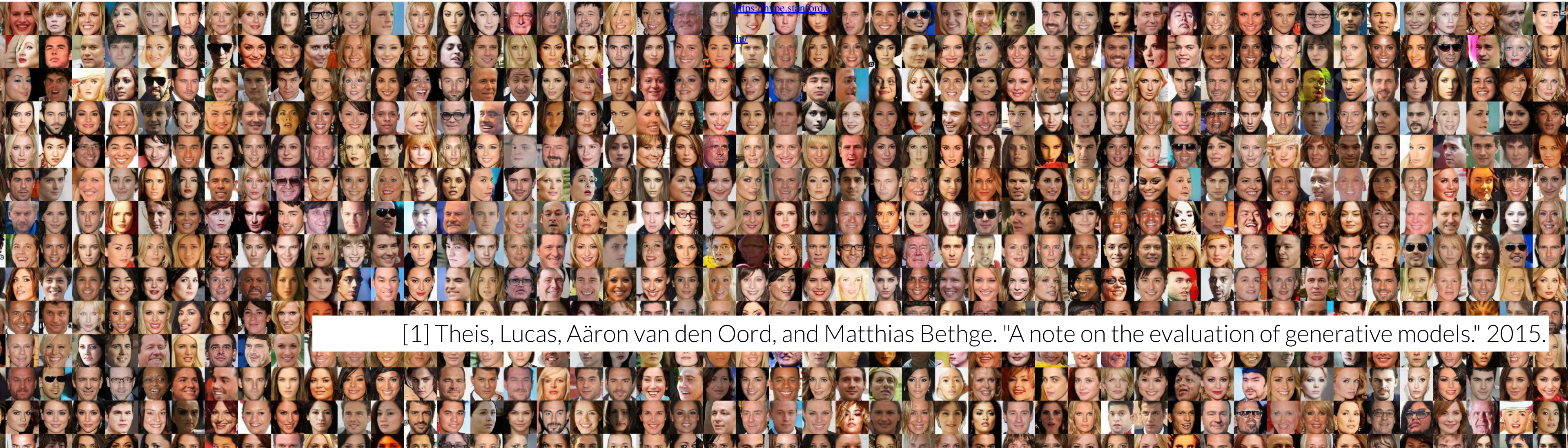
- Trained on imagenet
- **Inception score** is maximized when entropy of predicted output is low
    - Meaning if Inception says with high certainty that it's a "person", the score will be higher
- **FID** calculates distributions from activations of an Inception-v3 layer
- What is the problem with this approach?

Why not use automated metrics?

https://hype.stanford.edu/

Why not use automated metrics?

Density estimation has even been shown to be misleading [1].



[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.

Why not use automated metrics?

Density estimation has even been shown to be misleading [1].

Automated evaluation metrics on sampled outputs (Inception Score [2], FID [3], Precision [4], etc.) rely on ImageNet embeddings.

https://hype.stanford.edu/

[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.
[2] Salimans, Tim, et al. "Improved techniques for training GANs." 2016.
[3] Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." 2017.
[4] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." 2018.

APA

Why not use automated metrics? Or human metrics?

Density estimation has even been shown to be misleading [1].

Automated evaluation metrics on sampled outputs (Inception Score [2], FID [3], Precision [4], etc.) rely on ImageNet embeddings.

[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.
[2] Salimans, Tim, et al. "Improved techniques for training GANs." 2016.
[3] Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." 2017.
[4] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." 2018.

APA

Why not use automated metrics? Or human metrics?

Density estimation has even been shown to be misleading [1].

Automated evaluation metrics on sampled outputs (Inception Score [2], FID [3], Precision [4], etc.) rely on ImageNet embeddings.

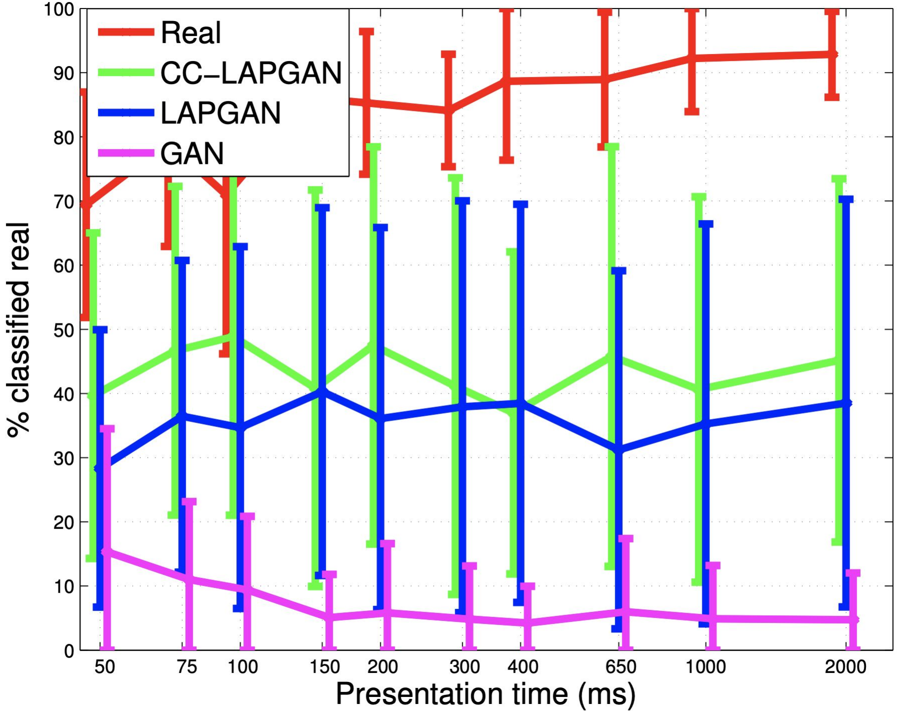
Human evaluation metric are ad-hoc — unreliable and costly.

[1] Theis, Lucas, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models." 2015.
[2] Salimans, Tim, et al. "Improved techniques for training GANs." 2016.
[3] Heusel, Martin, et al. "GANs trained by a two time-scale update rule converge to a local nash equilibrium." 2017.
[4] Sajjadi, Mehdi SM, et al. "Assessing generative models via precision and recall." 2018.

APA

Why not use human evaluation?

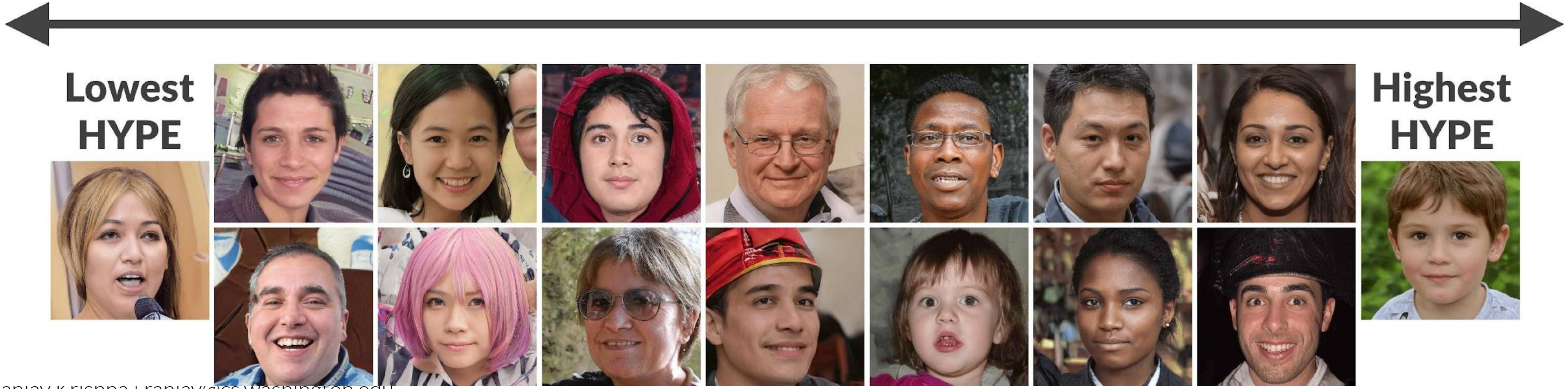
1. Ad-hoc, each executed in idiosyncrasy without proof of reliability or grounding to theory.
2. High variance in their estimates.
3. Lack clear separability between models.
4. Expensive and time-consuming



HYPE measures this progress using human evaluation that is consistent, efficient, and grounded in theory

HYPE is designed to address these problems:

1. **Grounded** method inspired by psychophysics methods in perceptual psychology.
2. **Reliable** and consistent estimator.
3. Statistically **separable** to enable a comparative ranking.
4. Cost and time **efficient**.



Psychophysics method: adaptive staircase procedure

- Staircase methods can determine human perceptual thresholds efficiently and reliably (Cornsweet, 1962).

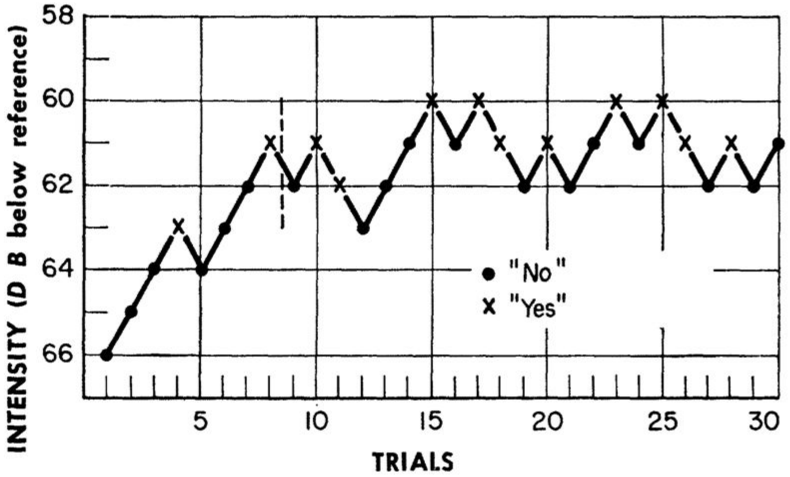
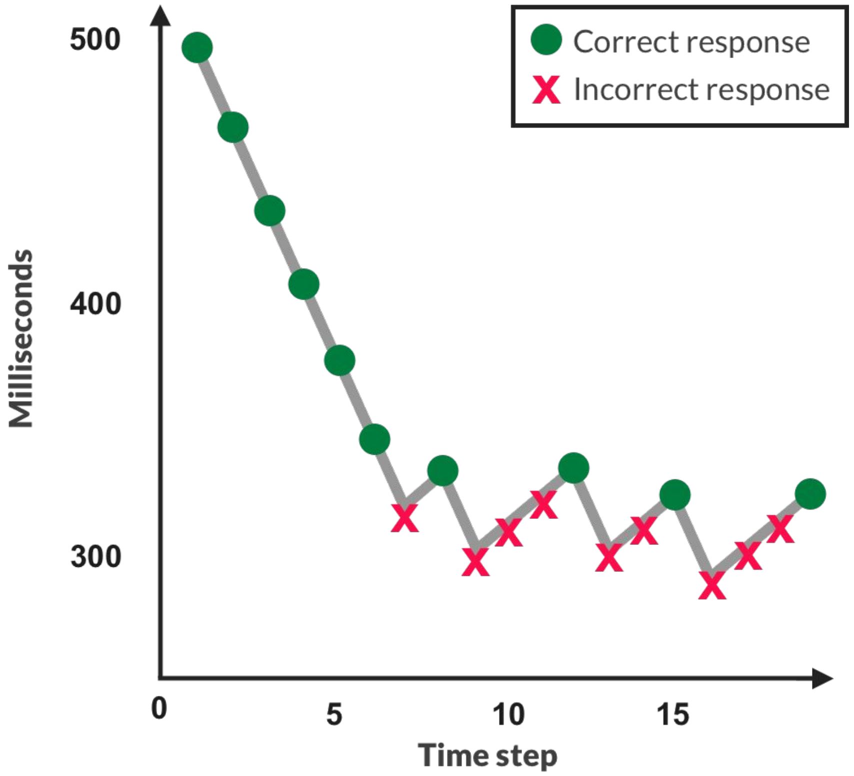


FIG. 1. DATA FROM THE DETERMINATION OF A TYPICAL AUDITORY THRESHOLD BY THE STAIRCASE-METHOD

HYPE: adaptive staircase procedure

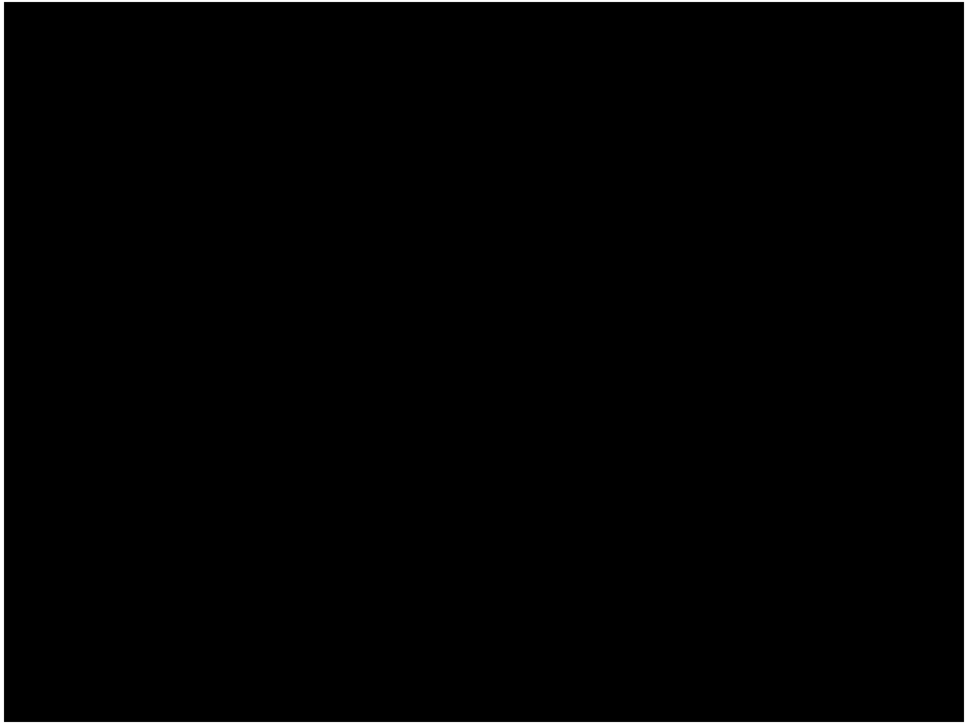


Time: 375ms

real

or

fake

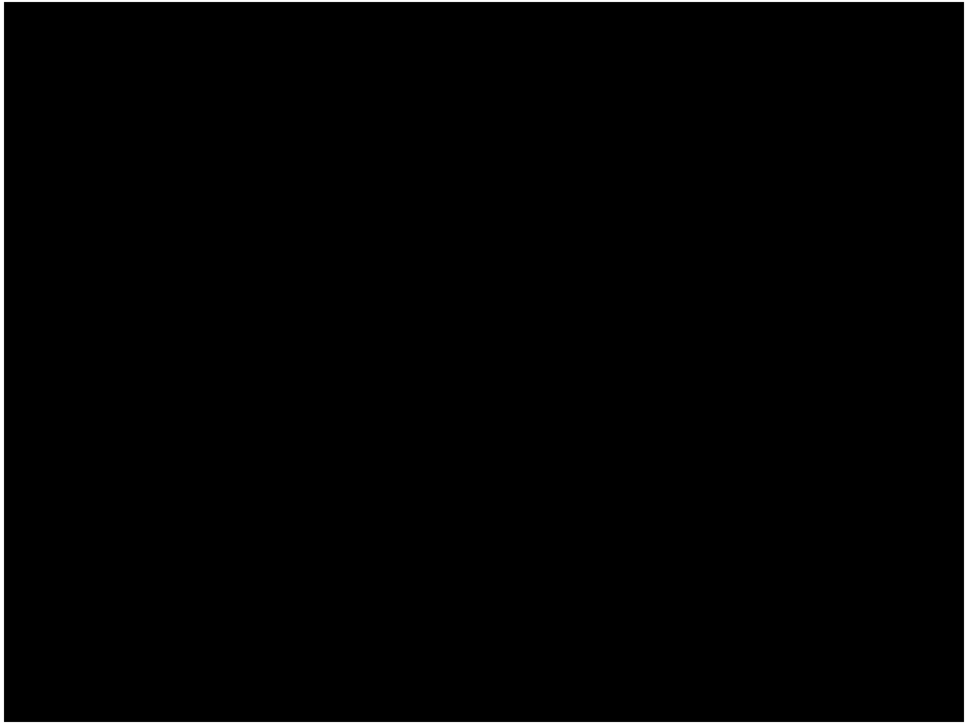


Time: 500ms

real

or

fake

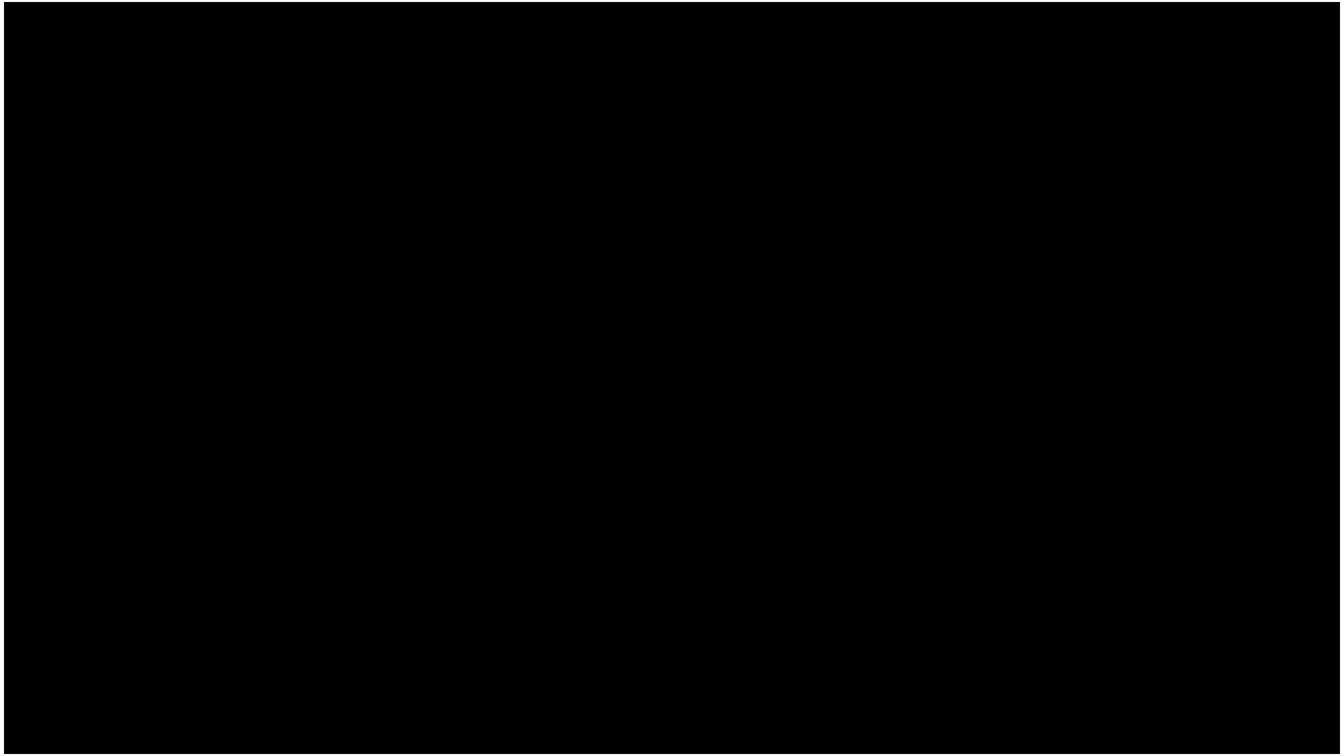


Time: 250ms

real

or

fake

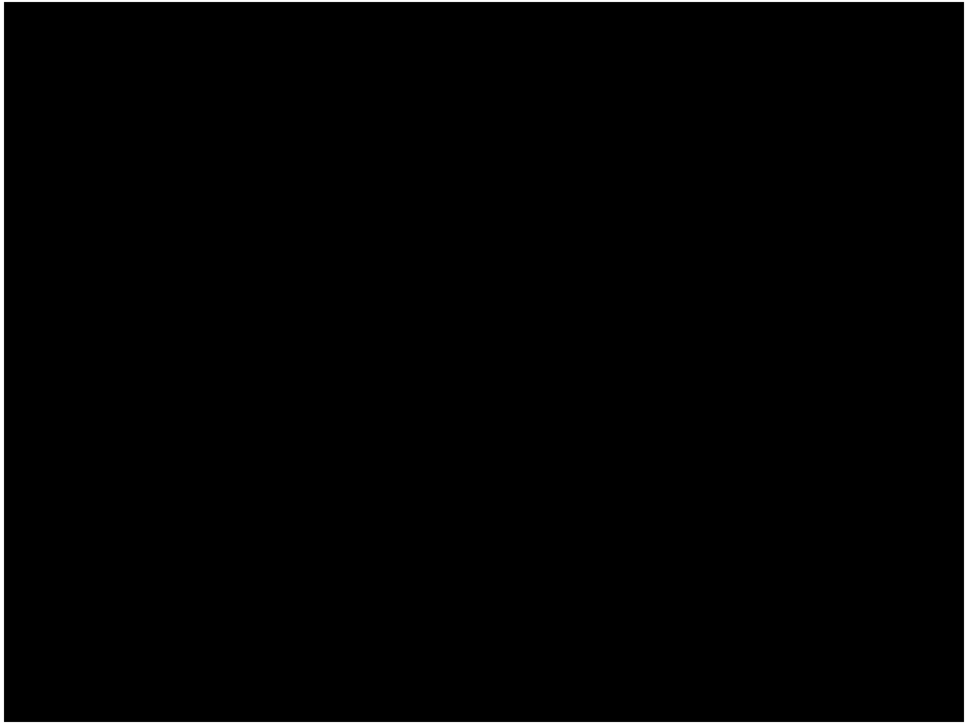


Time: 125ms

real

or

fake



Creating a reliable score

To ensure reliability, we need to:

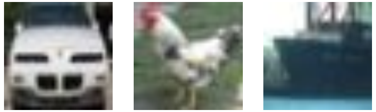
1. Hire and train/filter a sufficient number of evaluators.
2. Sample sufficient outputs.
3. Aggregate.

Experiments

Datasets

- CelebA

- CIFAR-10

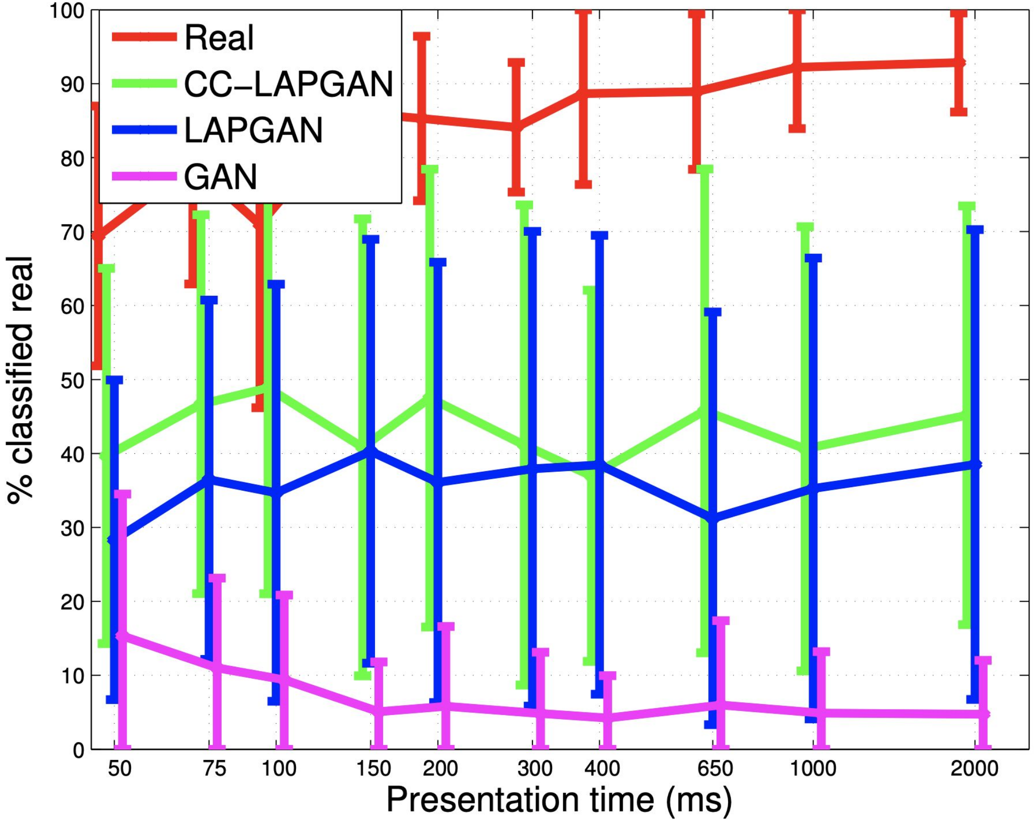


- ImageNet-5

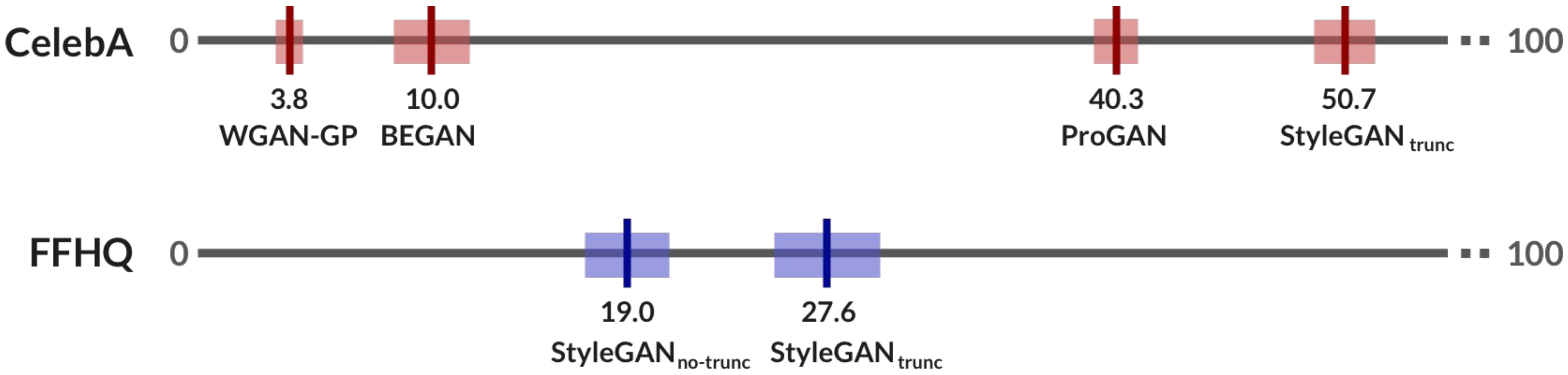
- FFHQ

Results

Are HYPE's results statistically separable?

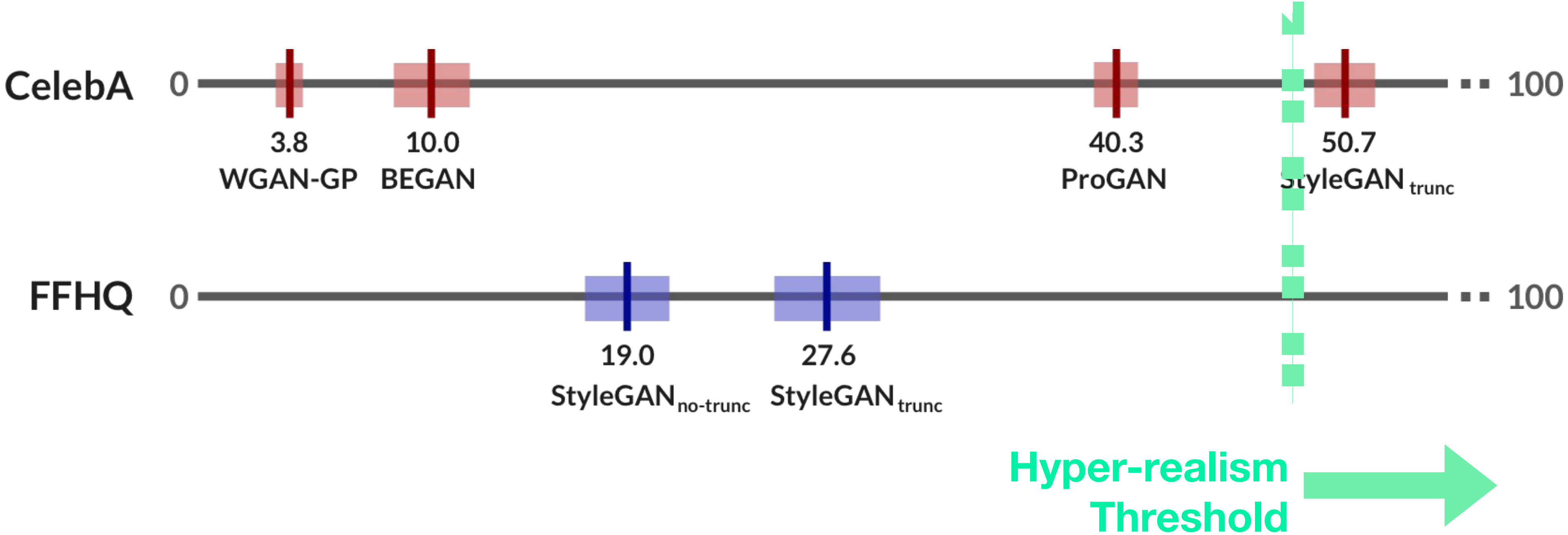


Are HYPE's results statistically separable?



Ranjay Krishna | ranjay@cs.washington.edu

Are HYPE's results statistically separable?



HYPE achieves:

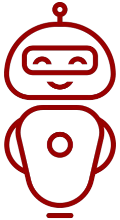
1. **Grounded** method inspired by psychophysics methods in perceptual psychology.
2. **Reliable** and consistent estimator.
3. Statistically **separable** to enable a comparative ranking.
4. Cost and time **efficient**.

Lecture 4

The challenges with understanding models

Ranjay Krishna | ranjay@cs.washington.edu

From evaluating AI to instead evaluating IA



Artificial Intelligence

Goal: Evaluate model generalization

Metrics: F1, accuracy, fairness, etc.

Can be automated

Human-Computer Interaction

Goal: Evaluate human task success

Metrics: Trust, correctness, interpretability, etc.

Often cannot be automated

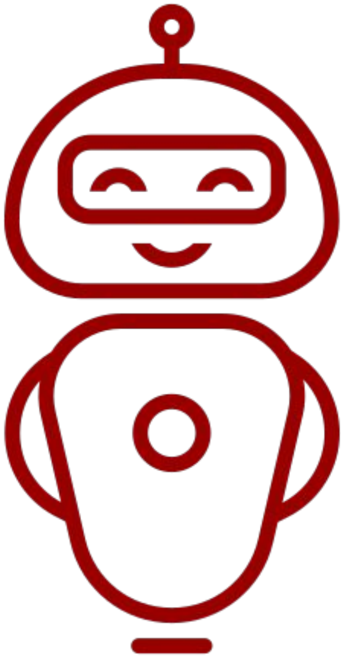
What does it mean to augment intelligence?

What does it really mean when people say human-centered AI?

- It's about dealing with users, with communities, and with societies
- It's a set of processes and guidelines through which we design AI.
- It's about serving human needs.

The old language of AI

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities



The old language of AI

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities

Teammates
Acts as a collaborator, interacts using language

The old language of AI

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities

Teammates
Acts as a collaborator, interacts using language

Assured autonomy
Sets goals, makes decisions, improves itself

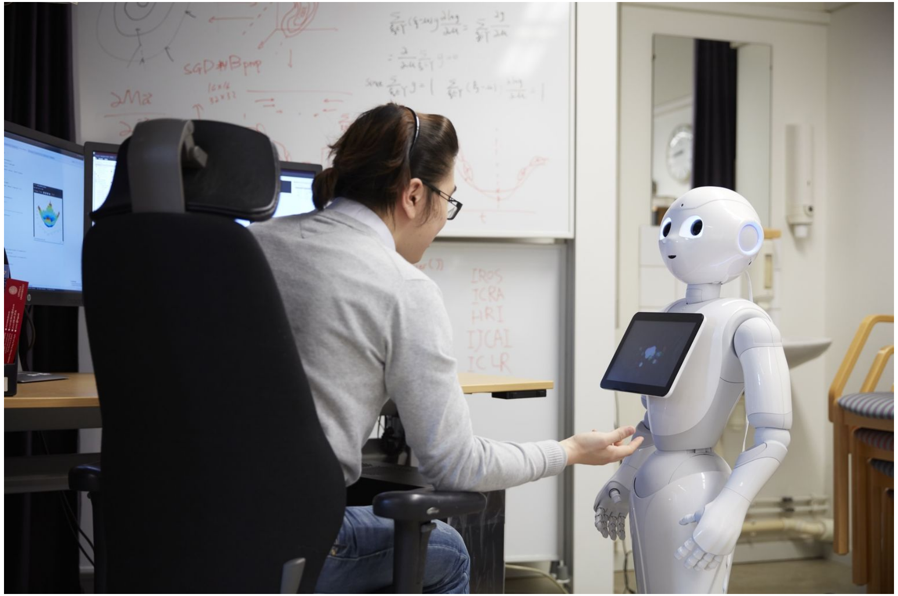
The old language of AI

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities

Teammates
Acts as a collaborator, interacts using language

Assured autonomy
Sets goals, makes decisions, improves itself

Social robots
Anthropomorphic, humanoid, emotionally intelligent



Reframing with new metaphors

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities

Teammates
Acts as a collaborator, interacts using language

Assured autonomy
Sets goals, makes decisions, improves itself

Social robots
Anthropomorphic, humanoid, emotionally intelligent

Reframing with new metaphors

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities

→ **Supertools**
Augments human abilities and performance

Teammates
Acts as a collaborator, interacts using language

Assured autonomy
Sets goals, makes decisions, improves itself

Social robots
Anthropomorphic, humanoid, emotionally intelligent

Reframing with new metaphors

Intelligent Agents
Manifests cognitive, linguistic, perceptual abilities

→

Supertools
Augments human abilities and performance

Teammates
Acts as a collaborator, interacts using language

→

Tele-bots
Boosts human perception & motor skills

Assured autonomy
Sets goals, makes decisions, improves itself

Social robots
Anthropomorphic, humanoid, emotionally intelligent

Reframing with new metaphors

Intelligent Agents Manifests cognitive, linguistic, perceptual abilities	⇨	**Supertools** Augments human abilities and performance
Teammates Acts as a collaborator, interacts using language	⇨	**Tele-bots** Boosts human perception & motor skills
Assured autonomy Sets goals, makes decisions, improves itself	⇨	**Control centers** Supports human control & situation awareness
Social robots Anthropomorphic, humanoid, emotionally intelligent		

Reframing with new metaphors

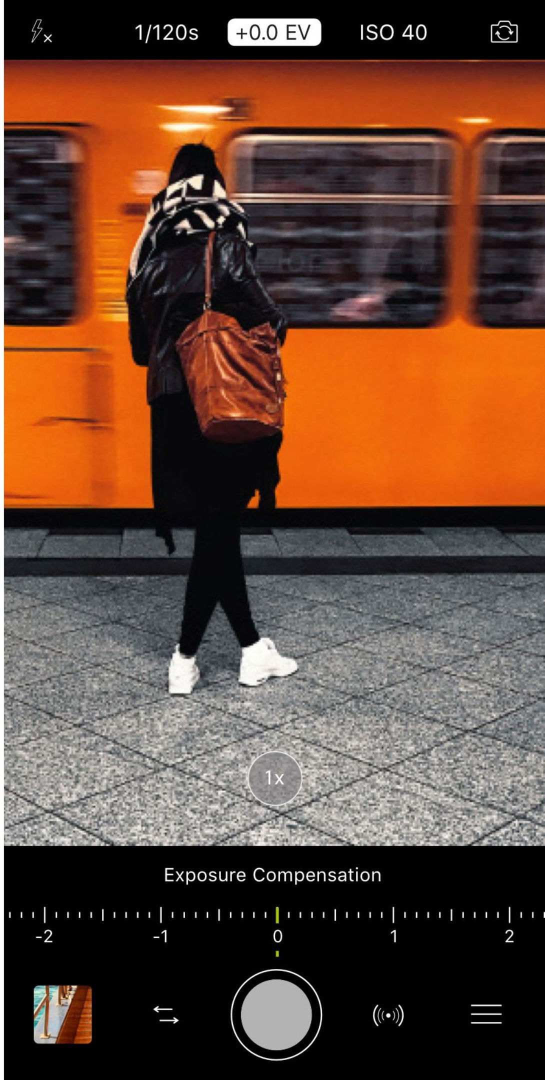
Intelligent Agents Manifests cognitive, linguistic, perceptual abilities	⇨	**Supertools** Augments human abilities and performance
Teammates Acts as a collaborator, interacts using language	⇨	**Tele-bots** Boosts human perception & motor skills
Assured autonomy Sets goals, makes decisions, improves itself	⇨	**Control centers** Supports human control & situation awareness
Social robots Anthropomorphic, humanoid, emotionally intelligent	⇨	**Active appliances** Low cost, easy to use, reliable applications

Putting these metaphors in context

- Color balances
- Corrects hand jitter
- Auto zoom
- Controls the shutter speed

But it augments humans:

- You frame it
- You compose it
- You decide how to share it



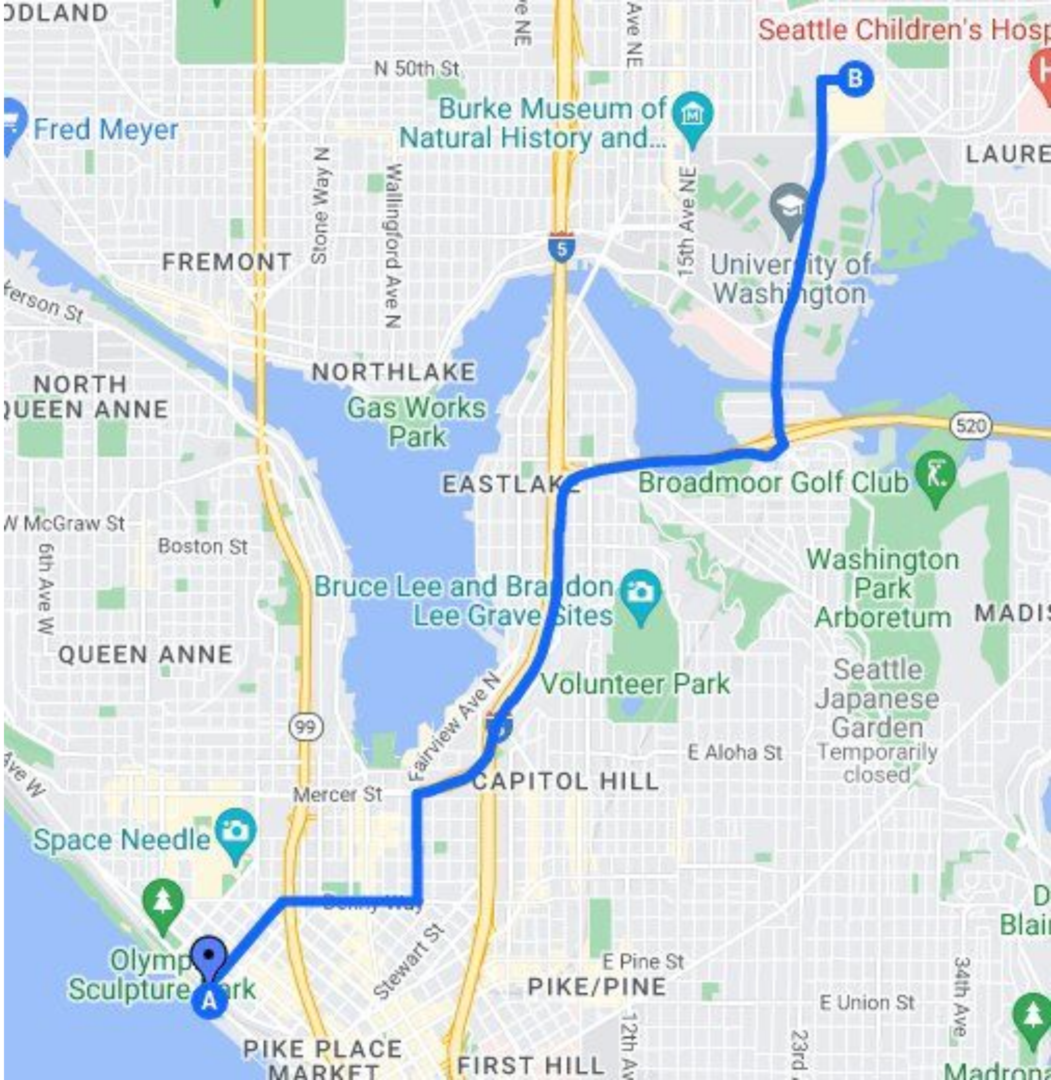
Another example

Lots of AI:

- Preview your route
- Get estimates of traffic

Augments you:

- You choose the route depending on your factors (optimal route, scenic route, gas needs, etc.)



Another example of a tele-bot

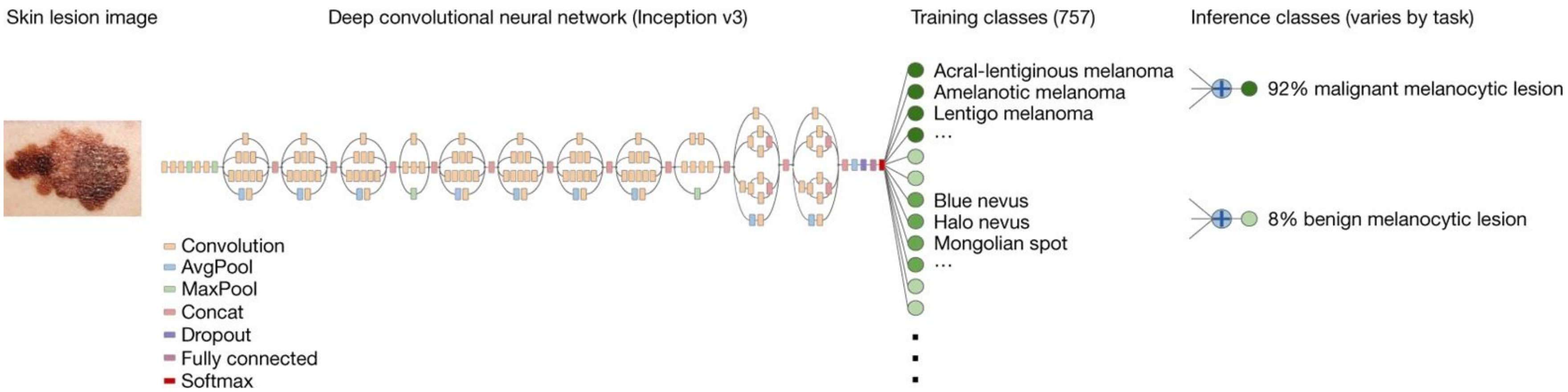
Da-vinci surgery bot

- Controlled by a human
- Augments human capabilities through with precision actions

Tons of AI bots that are active applications

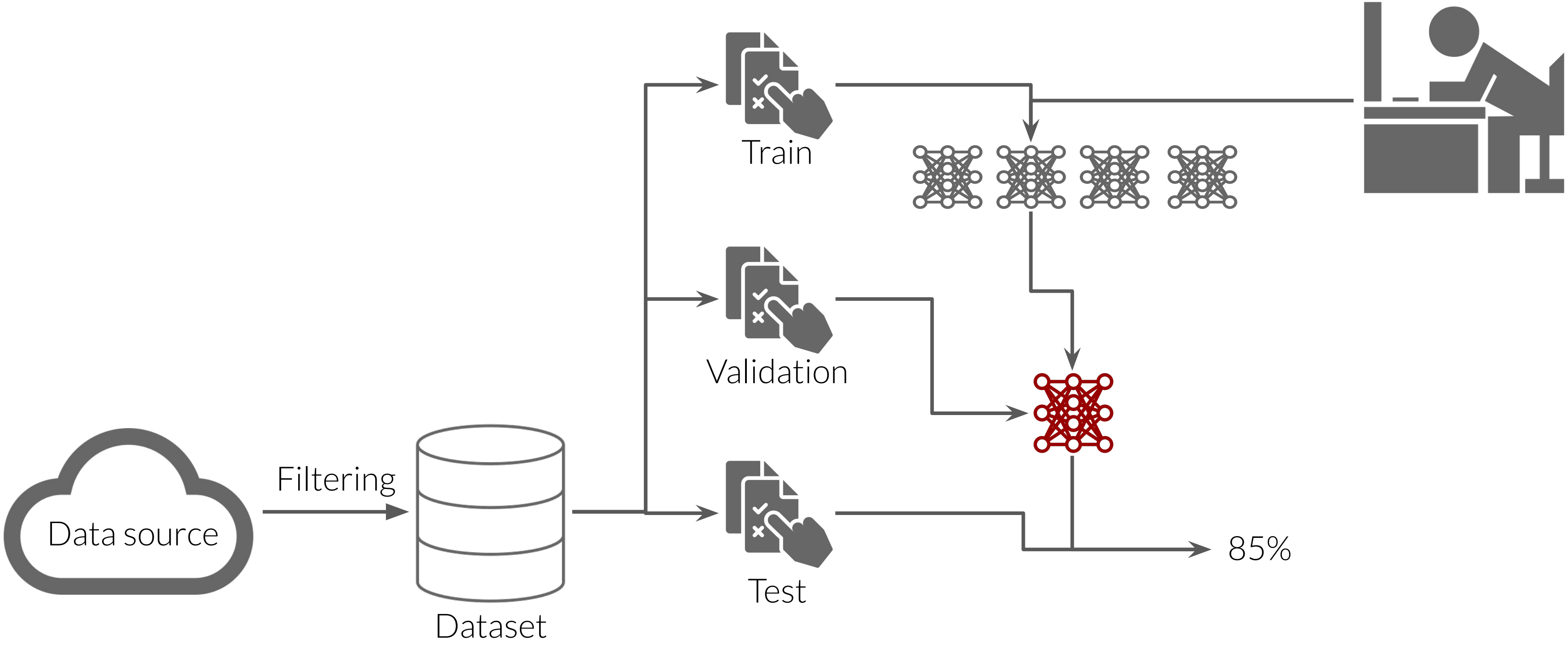
Ranjay Krishna | ranjay@cs.washington.edu

We are measuring model performance instead of human performance

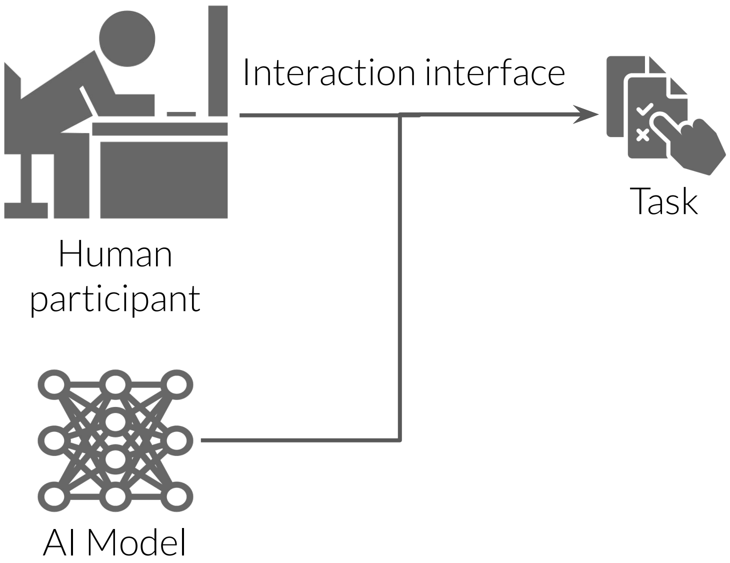


Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017

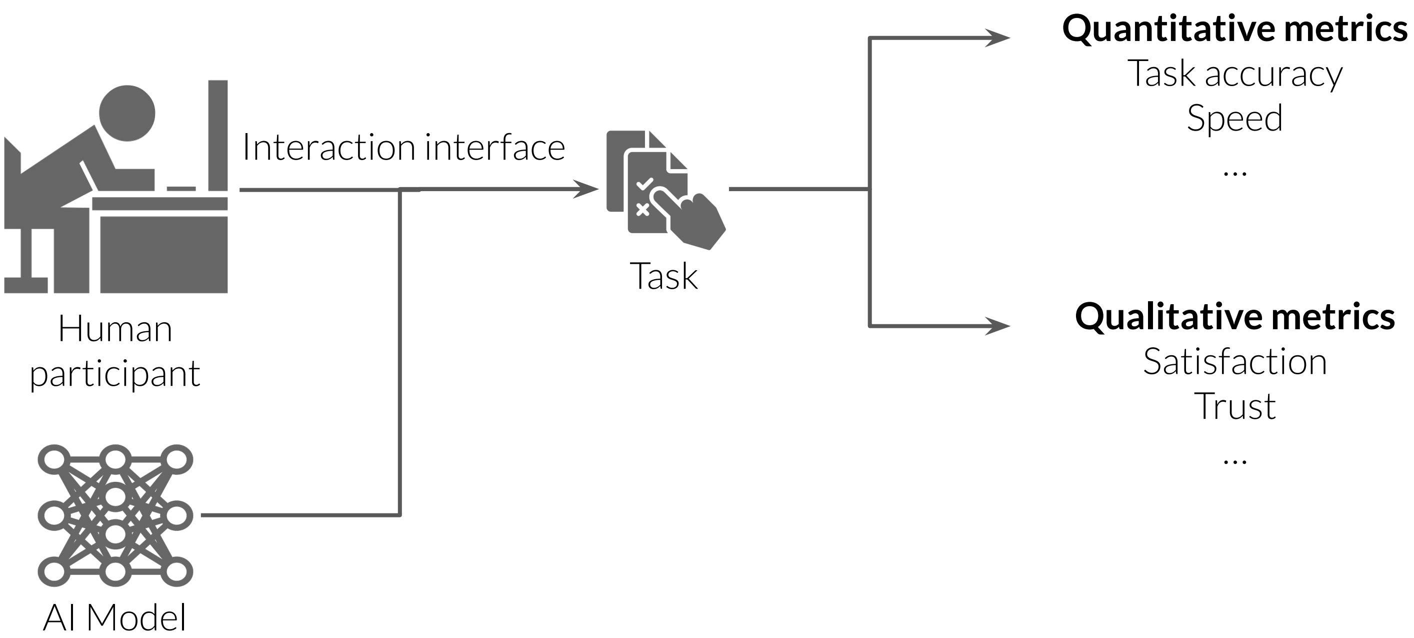
Last time: evaluation protocol for empirical machine learning



This time: evaluation protocol for human-AI systems



This time: evaluation protocol for human-AI systems



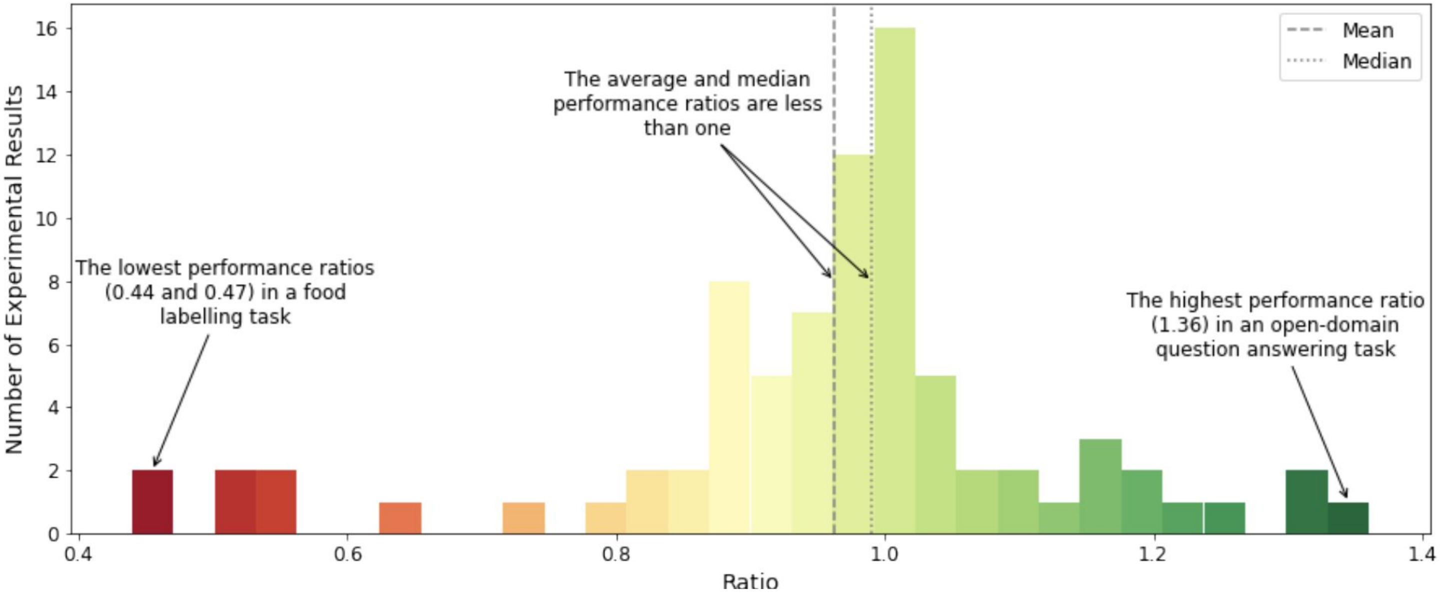
Human-AI teams ought to perform better but don't

$$\hat{\rho} = \frac{X_{HC}}{max(X_H, X_C)}$$

H = human

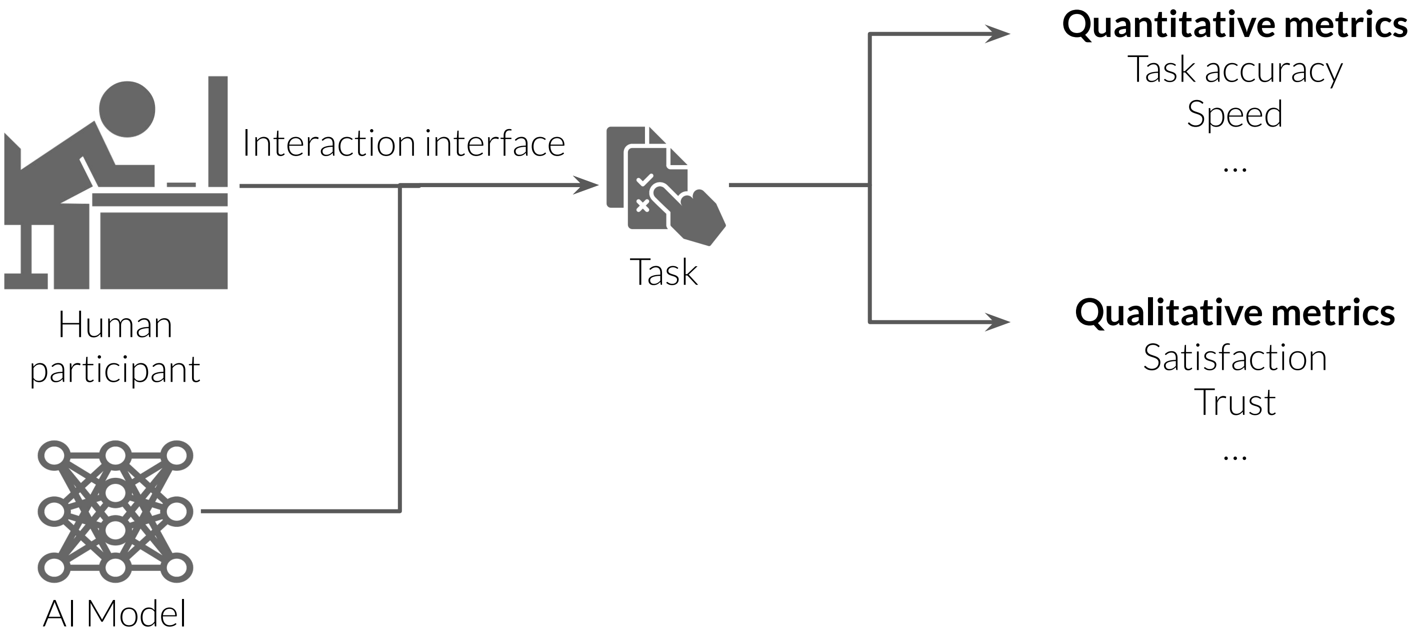
C = computer

HC = human-computer

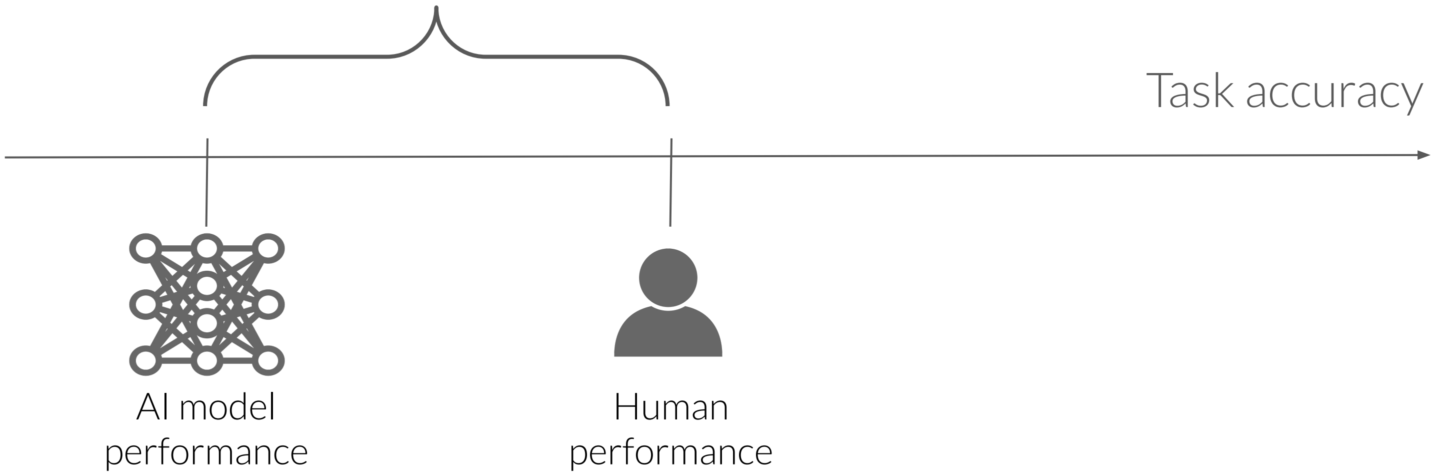


Campero et al. A test for evaluating performance in human-computer systems. ArXiv 2022

Class activity: What can go wrong with this setup?



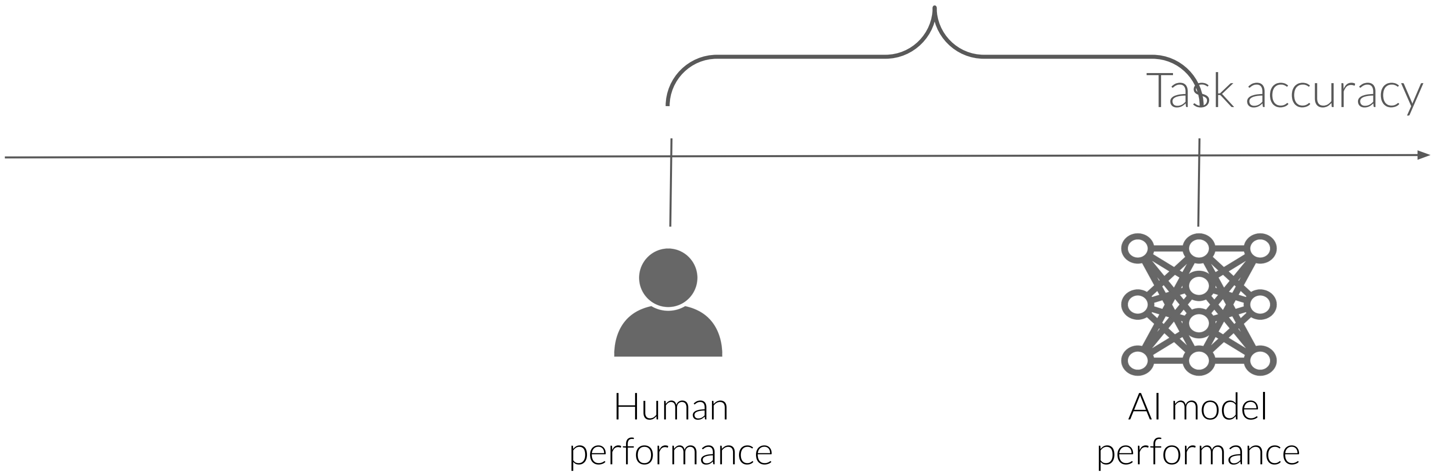
#1: Choice of AI model: useless if bad



Ranjay Krishna | ranjay@cs.washington.edu

Bansal et al. Does the Whole Exceed its Parts? The Efect of AI Explanations on Complementary Team Performance. CHI 2021

#1: Choice of AI model: Overrely if much better



Bansal et al. Does the Whole Exceed its Parts? The Efect of AI Explanations on Complementary Team Performance. CHI 2021

#2: Choice of metrics: Does it really measure human utility?

How do you think DALL-E evaluated their model?

this gray bird has a pointed beak black wings with small white bars long thigh and tarsus and a long tail relative to its size

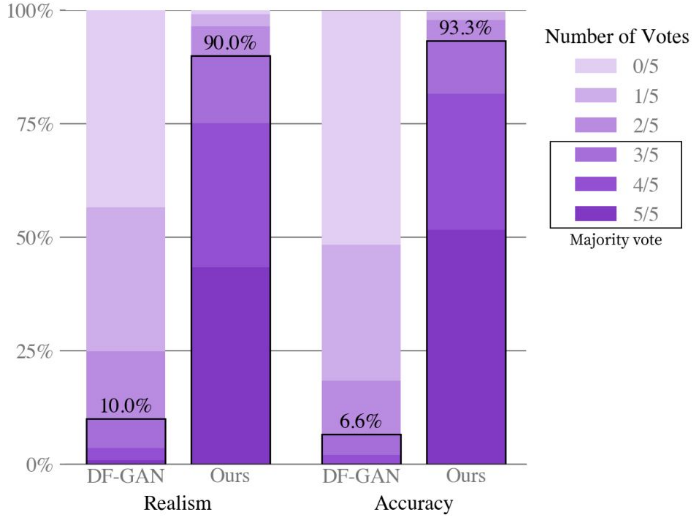
this rotund bird has a black tipped beak a black tail with a yellow tip and a black cheek patch

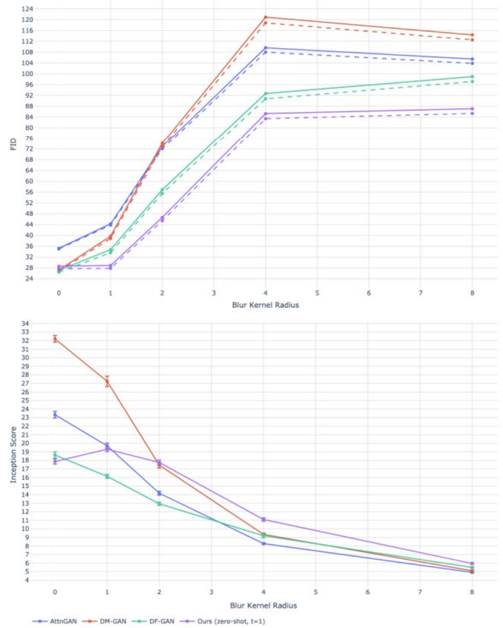
this is a small white bird with a yellow crown and a black eye ring and cheek patch and throat

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

#2: Choice of metrics: Does it really measure human utility?

How do you think DALL-E evaluated their model?



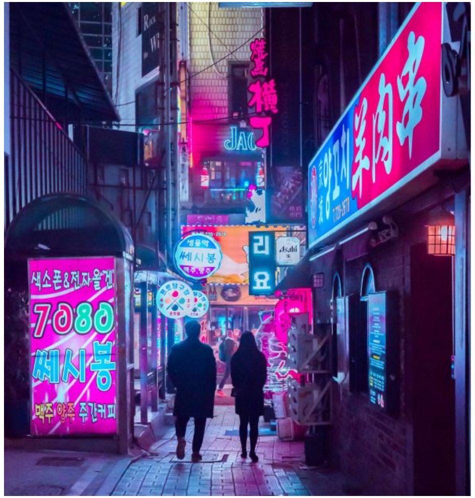


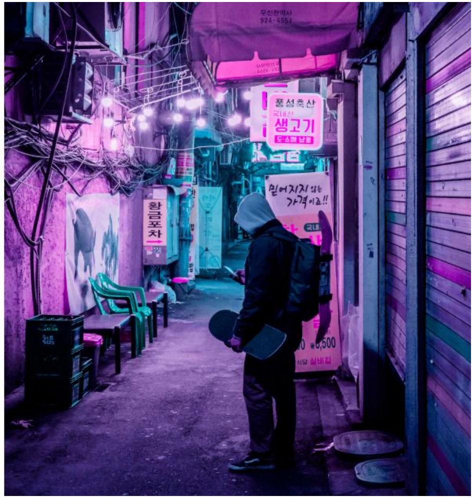
(a) FID and IS on MS-COCO as a function of blur radius.

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Does it really measure human utility?

Let’s try and generate some images similar to bladerunner scenes





Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Does it really measure human utility?

Seattle space needle with neon signage in the style of bladerunner

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Does it really measure human utility?

Seattle space needle with neon signage in the style of bladerunner

neon seattle space needle with **streets** in the style of bladerunner

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Does it really measure human utility?

Seattle space needle with neon signage in the style of bladerunner

neon seattle space needle with **streets** in the style of bladerunner

seattle space needle with **neon signs** and **nighttime rain** and **street market** in the style of bladerunner

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

After 18 iterations!!

Seattle space needle with neon signage in the style of bladerunner

neon seattle space needle with **streets** in the style of bladerunner

seattle space needle with **neon signs** and **nighttime rain** and **street market** in the style of bladerunner

Tall seattle space needle with neon signs and nighttime rain and street market and **people** in the style of bladerunner

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

After 18 iterations!!!

Seattle space needle with neon signage in the style of bladerunner

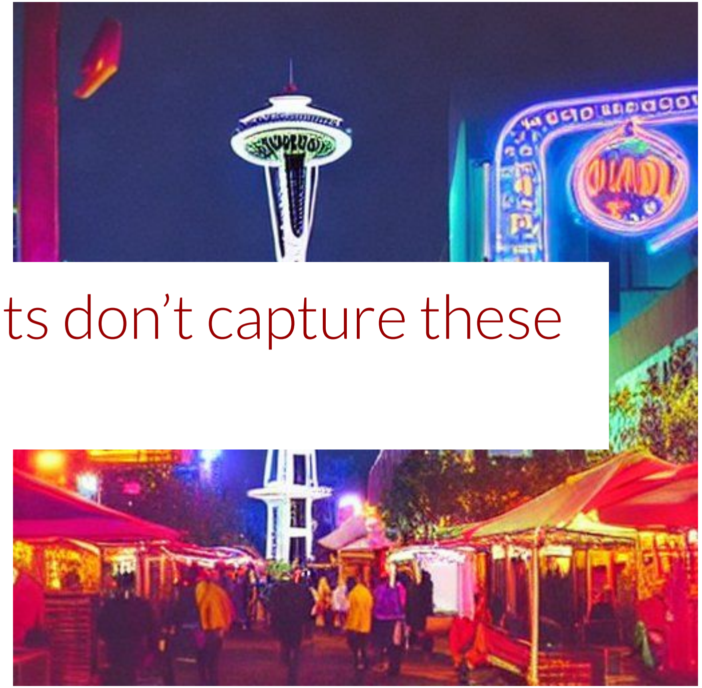
neon seattle space needle with streets in the style of bladerunner

seatt

rain ;

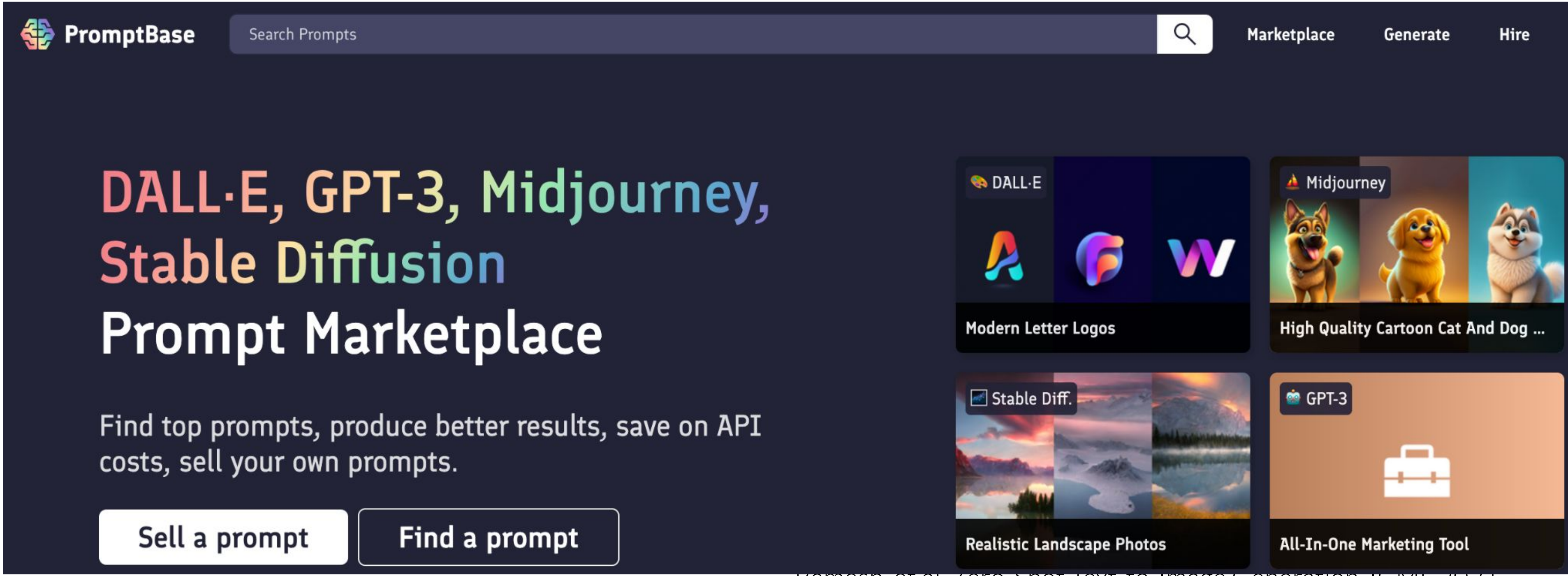
Realism and human judgements don’t capture these aspects of using the AI model

Tall seattle space needle with neon signs and nighttime rain and street market and **people** in the style of bladerunner



Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Prompt engineering is an unfortunate focus for many today but no way to evaluate their utility!



Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Language as an interaction modality

Seattle space needle with neon signage in the style of bladerunner

neon seattle space needle with **streets** in the style of bladerunner

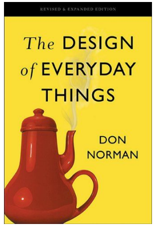
seattle space needle with **neon signs** and **nighttime rain** and **street market** in the style of bladerunner

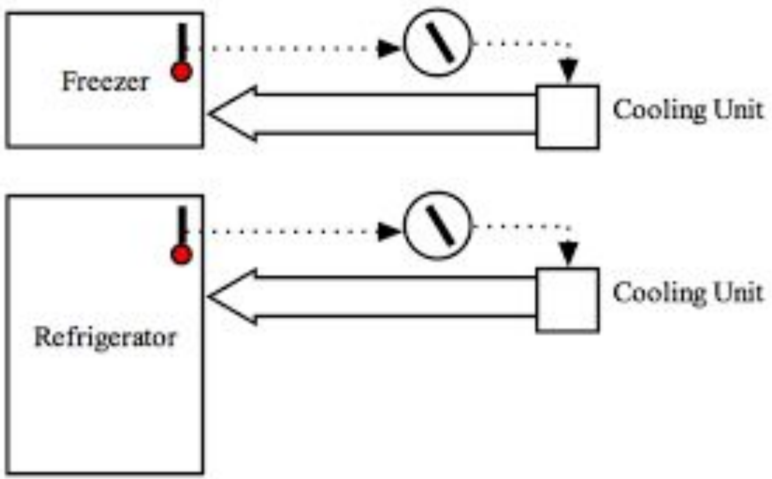
Tall seattle space needle with neon signs and nighttime rain and street market and **people** in the style of bladerunner

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Ranjay Krishna | ranjay@cs.washington.edu

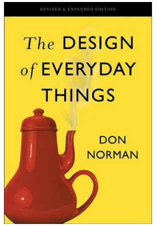
Grounding interactions to our conceptual models

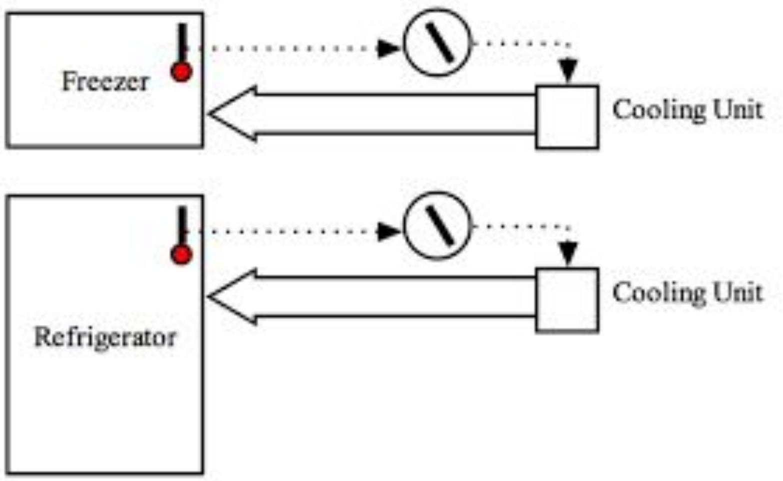




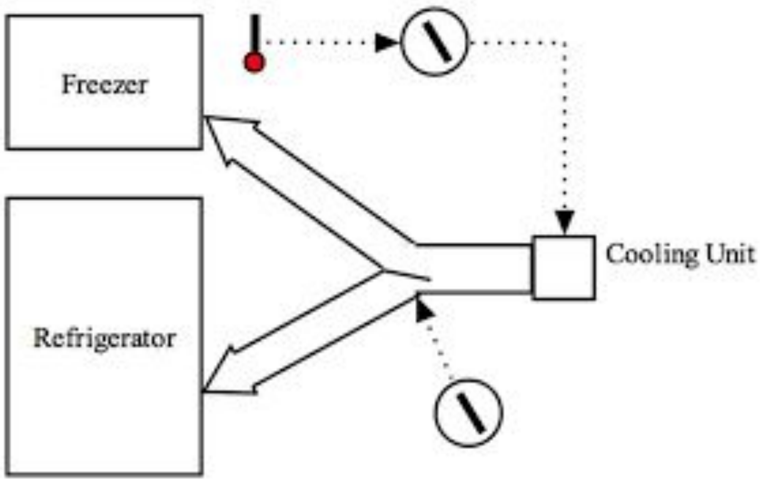
Two control units = two separate temperature controls

#3: Choice of interaction: Grounding interactions to our conceptual models





Two control units = two separate temperature controls



The real conceptual model

What conceptual model does this language interaction afford?

Seattle space needle with neon signage in the style of bladerunner

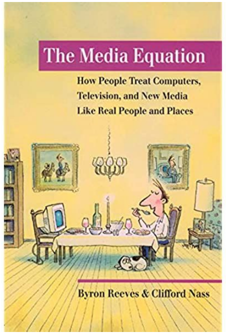
neon seattle space needle with **streets** in the style of bladerunner

seattle space needle with **neon signs** and **nighttime rain** and **street market** in the style of bladerunner

Tall seattle space needle with neon signs and nighttime rain and street market and **people** in the style of bladerunner

Ramesh et al. Zero-Shot Text-to-Image Generation. ICML 2021

Why language language interactions are appealing?

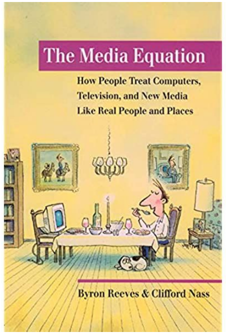


General communication theory:

- people assign human characteristics to computers, AI models, and other media to treat them as social actors.

- The thought process might go: *If people already treat machines as social actors, let's enable them to interact with language*

Why language language interactions are appealing?



More nuanced understanding of the media equation: when machines project social competence or enable social interactions, they induce shortcut social scripts in people

- In other words, when you allow people to interact with machines with language, they expect machines to competently react like people do

- The thought process might now go: *if I allow my model to interact with language, it should be able to do everything people can do with language: maintain context, repair through multiple interactions, explain its behavior, correct itself, ask for clarifications, ….*

Non-humans as teammates

- Police dogs and search and rescue dogs have a single handler.
- Incorporating them as equal teammates has failed

"Without self-interest and humanlike mental models, the introduction of a robot into a human team makes violations of trust and the ensuing consequences highly likely"

Groom and Nass. Can robots be teammates?. Interaction Studies 2007

#4: Choice of interface: The effects of anthropomorphisation

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

#4: Choice of interface: The effects of anthropomorphisation

Research question:

How do the words we use to describe an AI model change how people interact with them?

Khadpe et al. Conceptual Metaphors Affect Human-AI Collaboration. CSCW 2020

Conceptual Metaphors

Explains what a system might be capable of

A metaphor communicates expectations of what can and cannot be done with an AI model

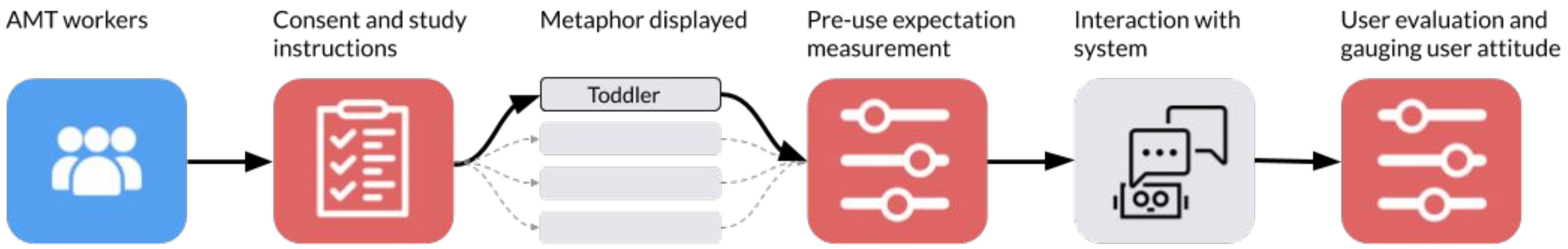
Visual Metaphors:


Audio Metaphors:

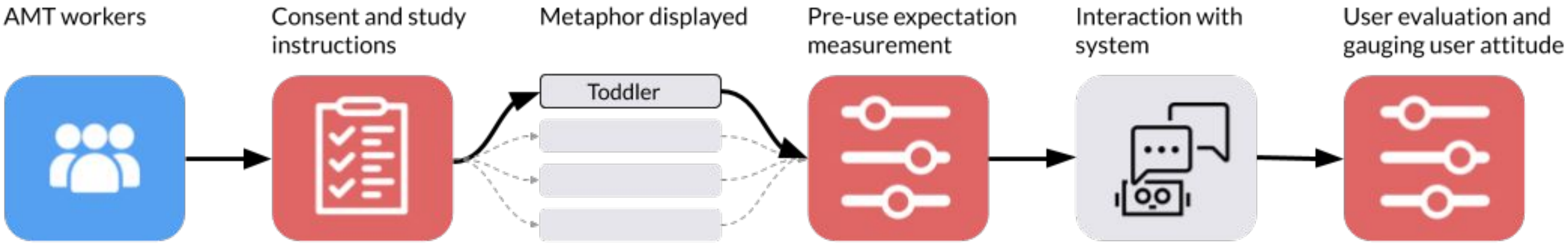
- Analog shutter clicking sound for mobile cameras

Textual Metaphors:

an administrative assistant, a teenager, a friend, or a psychotherapist



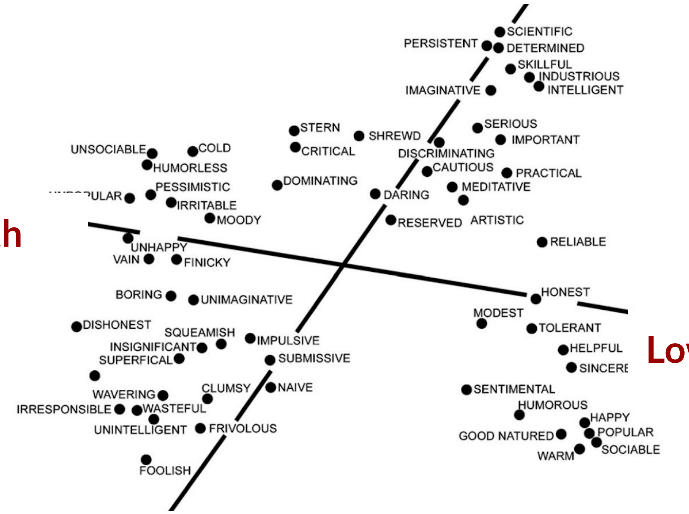
Study Workflow



How do you choose the metaphors?

High competence

Low warmth



Low warmth

Low competence

Fiske et al. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition . In Social cognition. 2018

How do conceptual metaphors impact evaluations?

Hypothesis 1: Based on the Assimilation Theory - people adapt experiences to match expectations

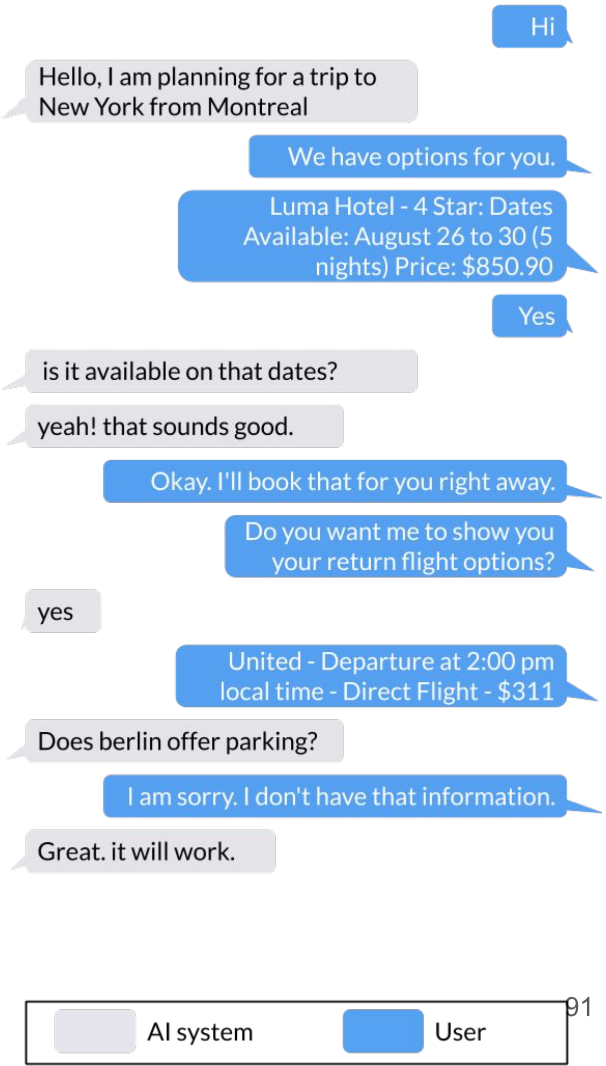
Positive metaphors (high competence, high warmth) -> positive evaluations

Hypothesis 2: Based on the Contrast Theory - people are attuned to a difference between expectations and experiences

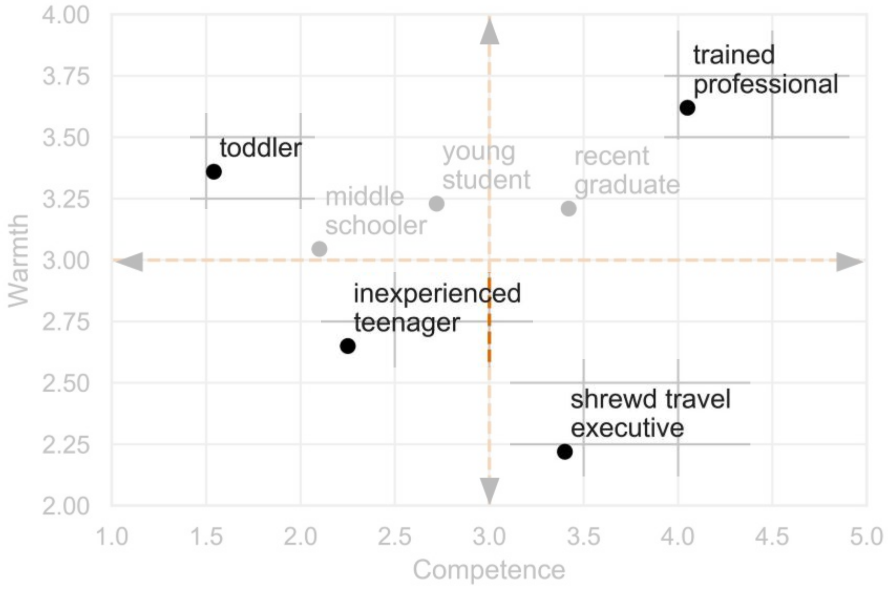
Positive metaphors (high competence, high warmth) -> poor evaluations

Muzafer Sherif, Daniel Taub, and Carl I Hovland. 1958. Assimilation and contrast effects of anchoring stimuli on judgments. Journal of experimental psychology 55, 2 (1958), 150.

Wizard of Oz task for booking hotel, flights



We sampled metaphors along these two dimensions

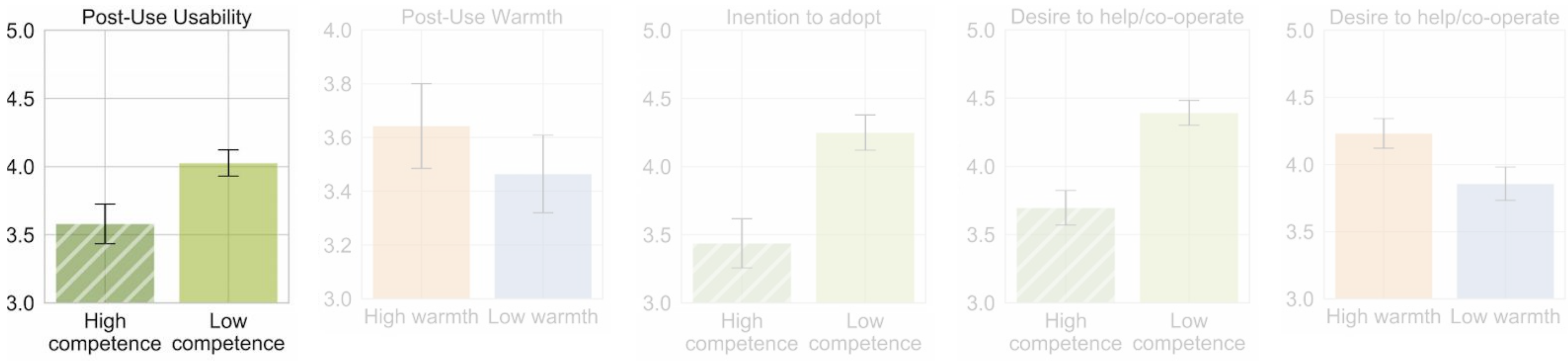


Manipulations:
4 treatment Groups + 1 Control Group that is not shown a metaphor

Class guesses: What do you think happens?

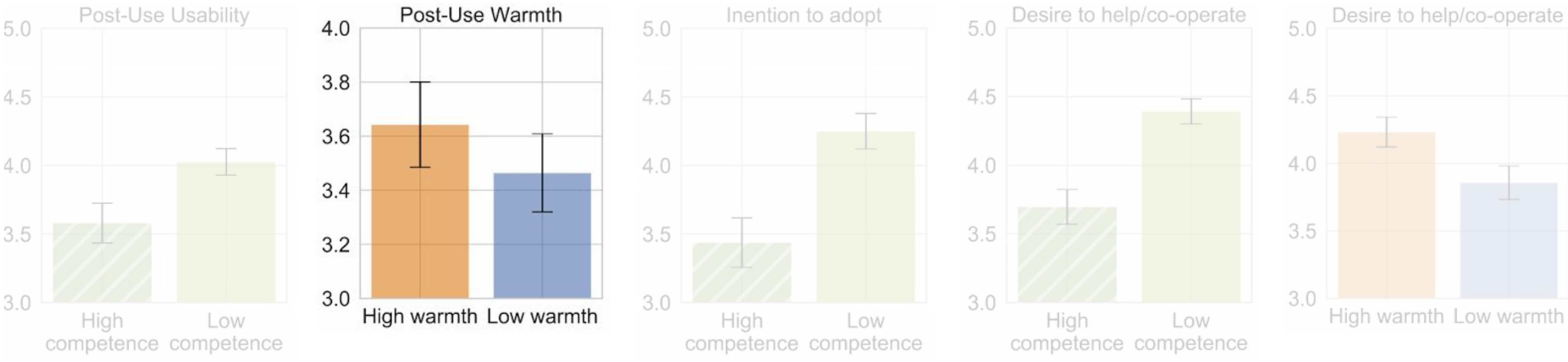
Variables we manipulate		Variables we measure
		Do you think the AI is **usable**?
competence		Do you think the AI is **Warm**?
	→	Will you **adopt** this AI?
Warmth		Would you **cooperate** with this AI model?

Users perceive agents with low competence to be more usable

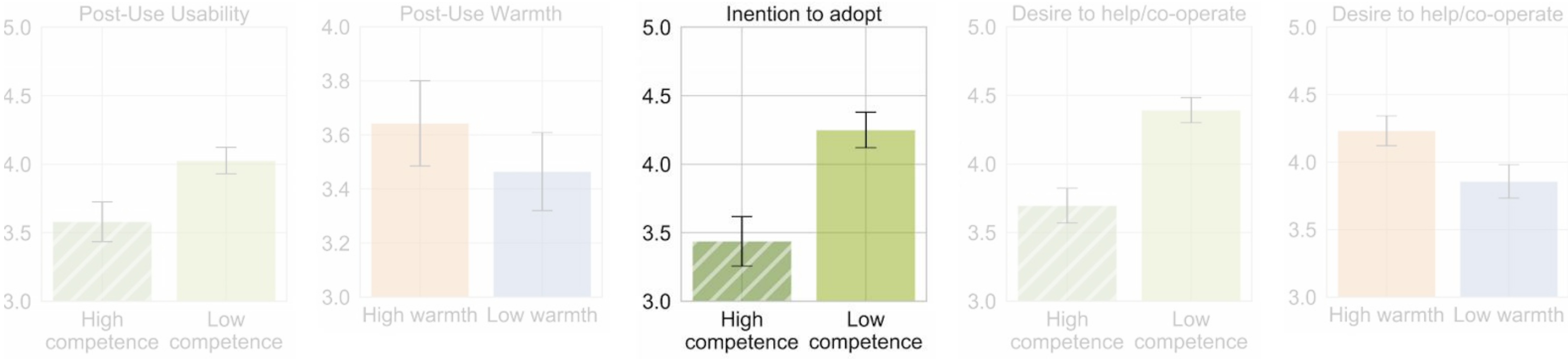


- Support for H2 and Contrast Theory - over performing expectations leads to positive evaluations

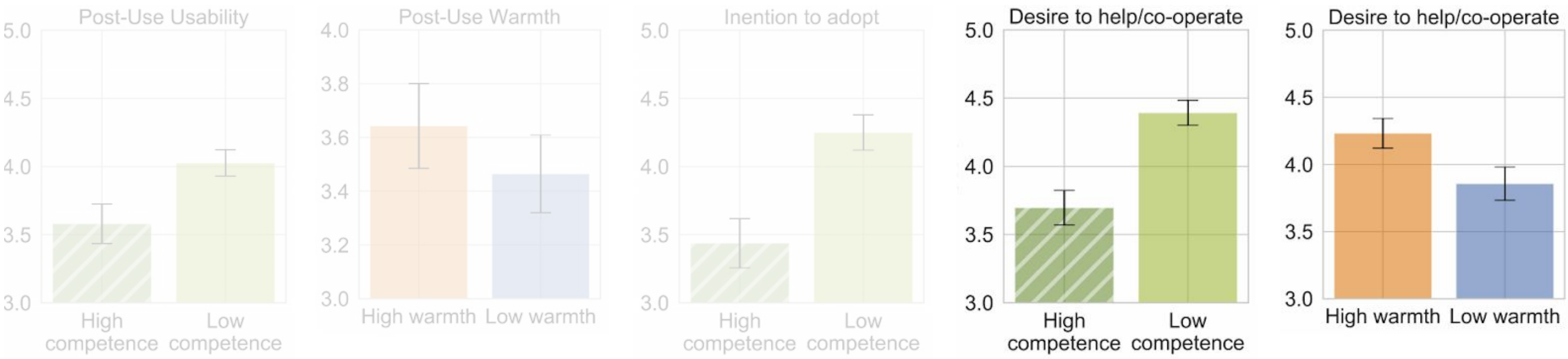
Metaphors directly affect how warm people think models are



Low competence metaphors increase users' likelihood of adopting the AI agent

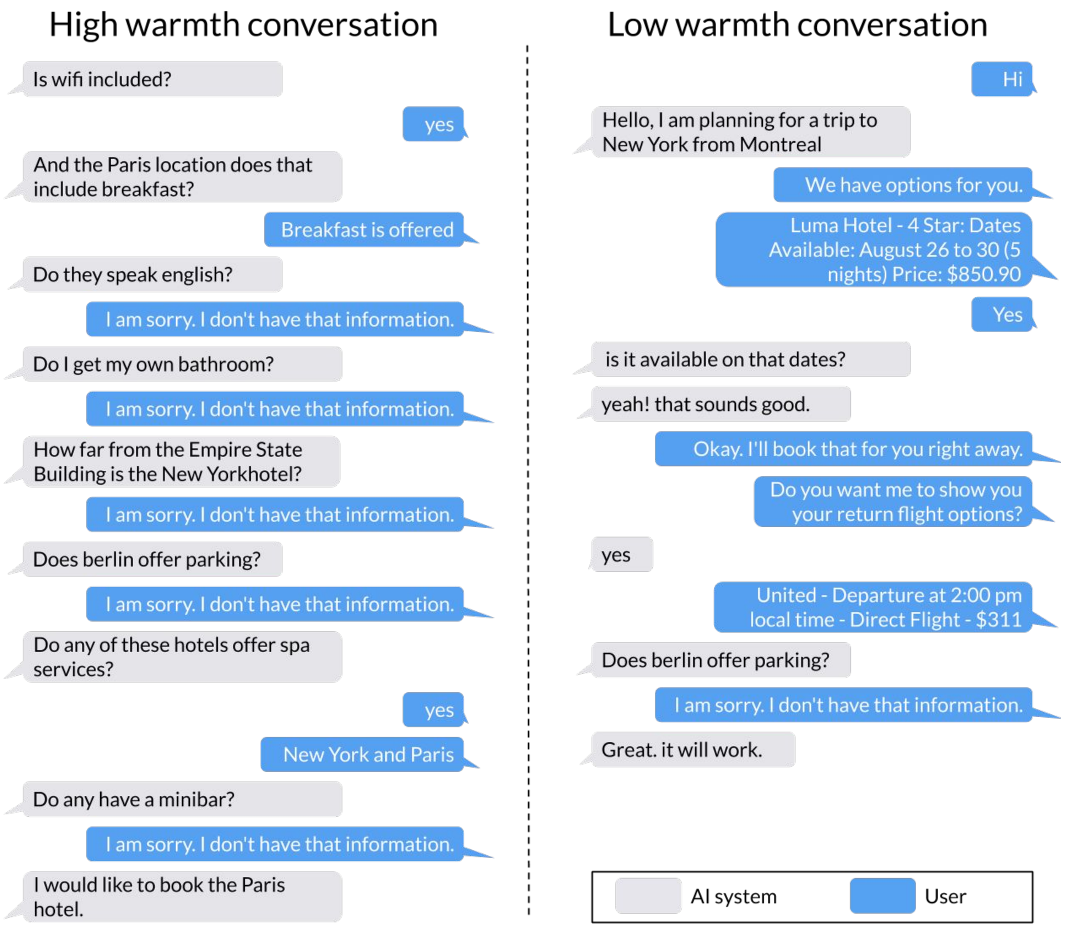


Users prefer to cooperate with agents that have high warmth and low competence

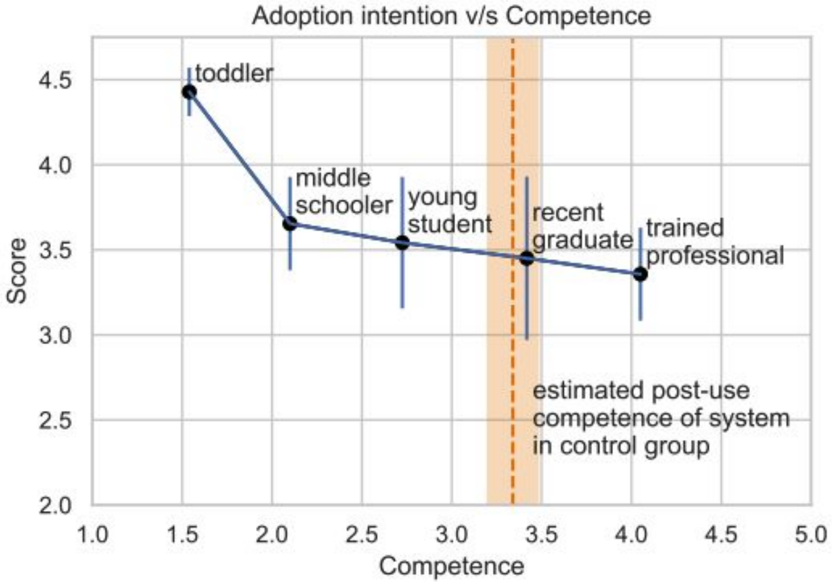


- mixed support to both H1 and H2:
  - assimilation theory along the warmth dimension
  - contrast theory along the competence dimension.

Users use more words and spend more time speaking to agents with high warmth

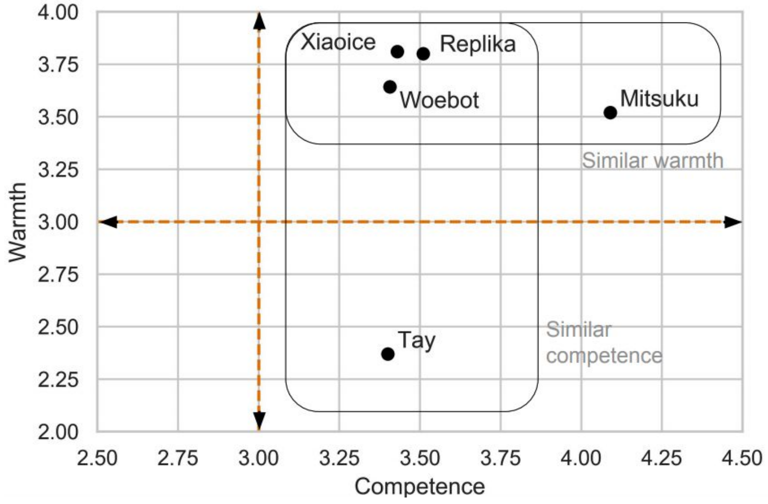


effect is greater as the violation is greater



Extreme violations of expectations have stronger effects

Retrospective Analysis



Most chabots today signal high competence.
=> users are left disappointed

Xiaoice is seen as having higher warmth as Tay, which could explain why Tay was subject to a lot more antisocial behaviour

Similarly Woebot and Replica are high warmth and elicit positive behaviour .

Mitsuku is seen as high competence which could explain it's dehumanisation

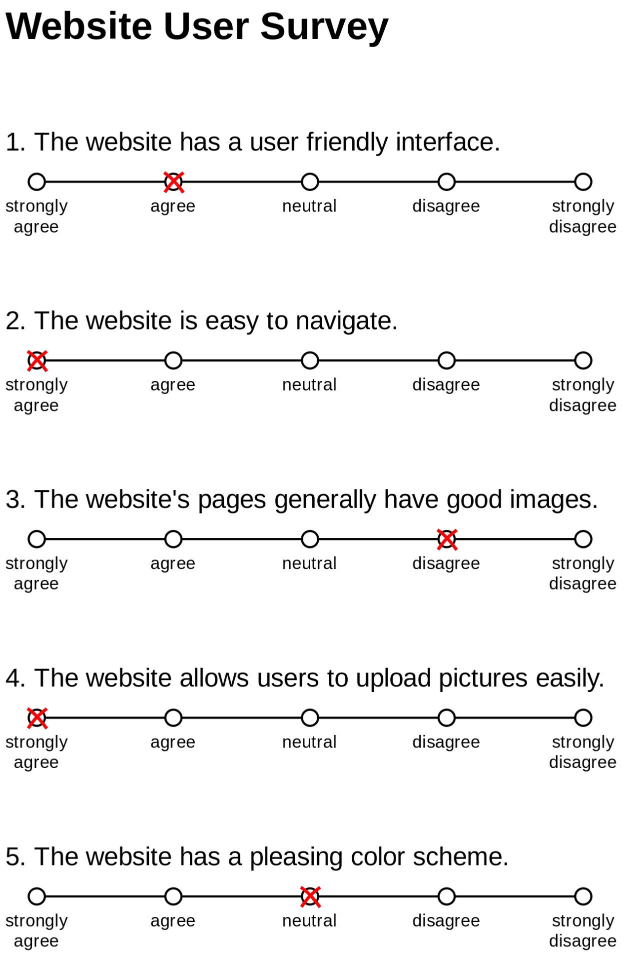
#5: Choice of aggregation:

Subjective interpretations violate absolute values

Linear assumption violates normalization

Averaging across participants doesn't work

Paper suggests asking people to guess with what probability they prefer X over Y. And Y over X.



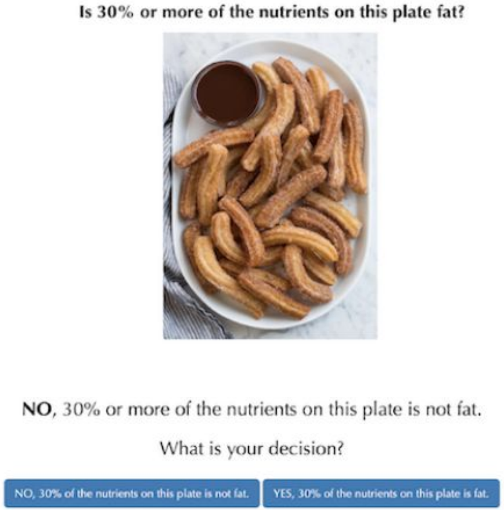
Ethayarajh et al. The Authenticity Gap in Human Evaluation. ArXiv 2022

#6: Choice of task: Proxy task (left) doesn't correlate with actual task (right)

The actual task:

- Is there >30% fat?

AI predicts binary (yes/no) answer



Bucinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020

#6: Choice of task: Proxy task (left) doesn't correlate with actual task (right)

The actual task:

- Is there >30% fat?

AI predicts binary (yes/no) answer

AI can produce explanations in the form of exemplars.

Bucinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020

#6: Choice of task: Proxy task (left) doesn't correlate with actual task (right)

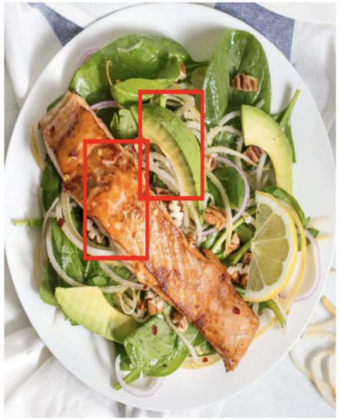
The actual task:

- Is there >30% fat?

AI predicts binary (yes/no) answer

AI can produce explanations in the form of detected concepts.

Is 30% or more of the nutrients on this plate fat?



Here are ingredients the AI recognized as main nutrients which make up 30% or more fat on this plate:

salmon
avocado

This AI recommended answer is:

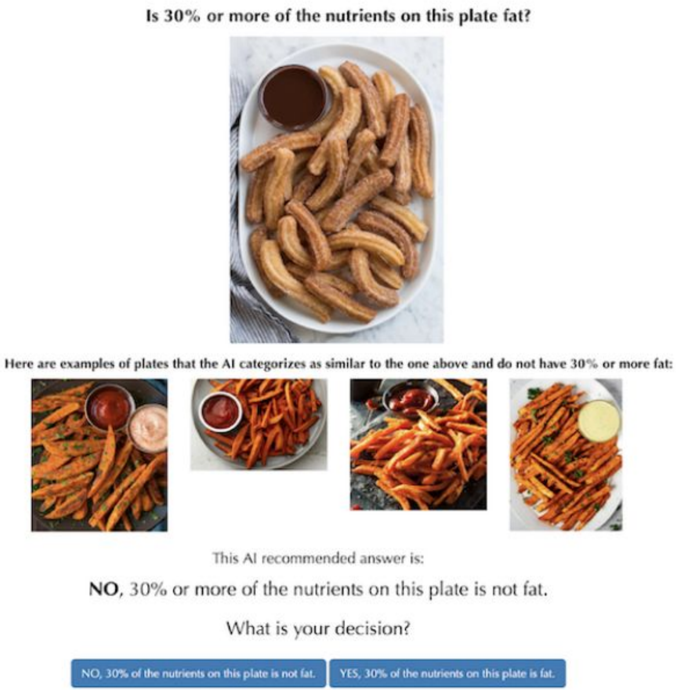
YES, 30% or more of the nutrients on this plate is fat.

What is your decision?

NO, 30% of the nutrients on this plate is not fat. | YES, 30% of the nutrients on this plate is fat.

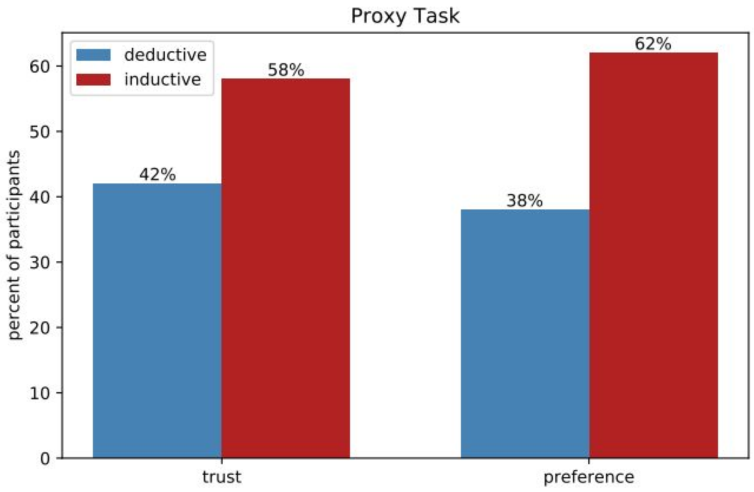
Bucinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020

The proxy task: What do you think the AI will choose?



Bucinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020

#6: Choice of task: Proxy tasks don't correlate with actual task



Deductive explanations = detected concepts

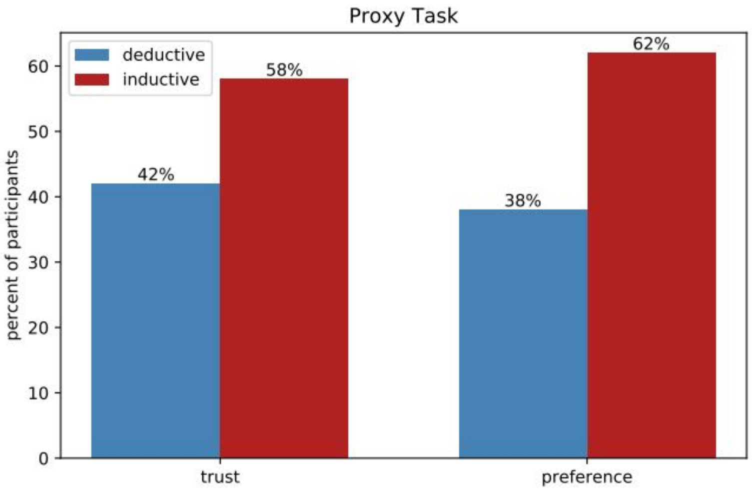
Use that information to deduce the answer

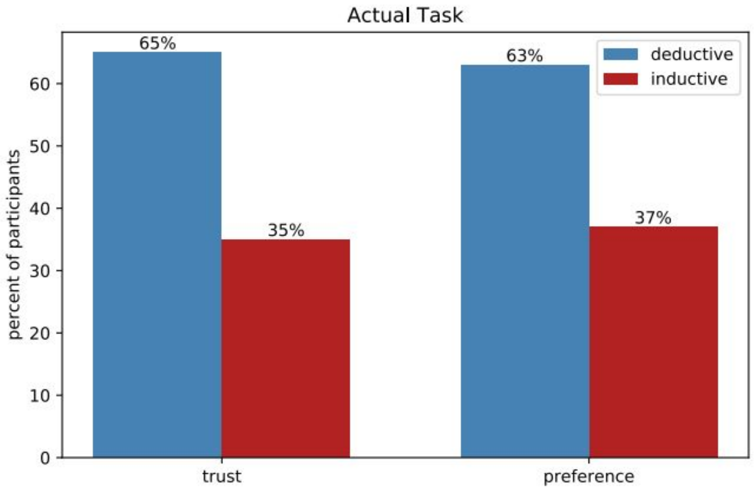
Inductive explanations: examplars

Use general patterns from other examples

Bucinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020

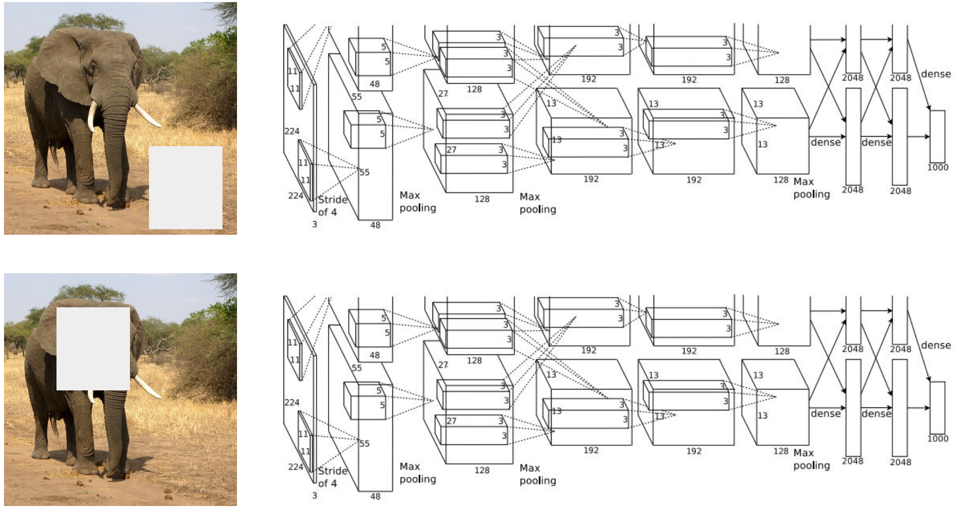
#6: Choice of task: Proxy tasks don't correlate with actual task

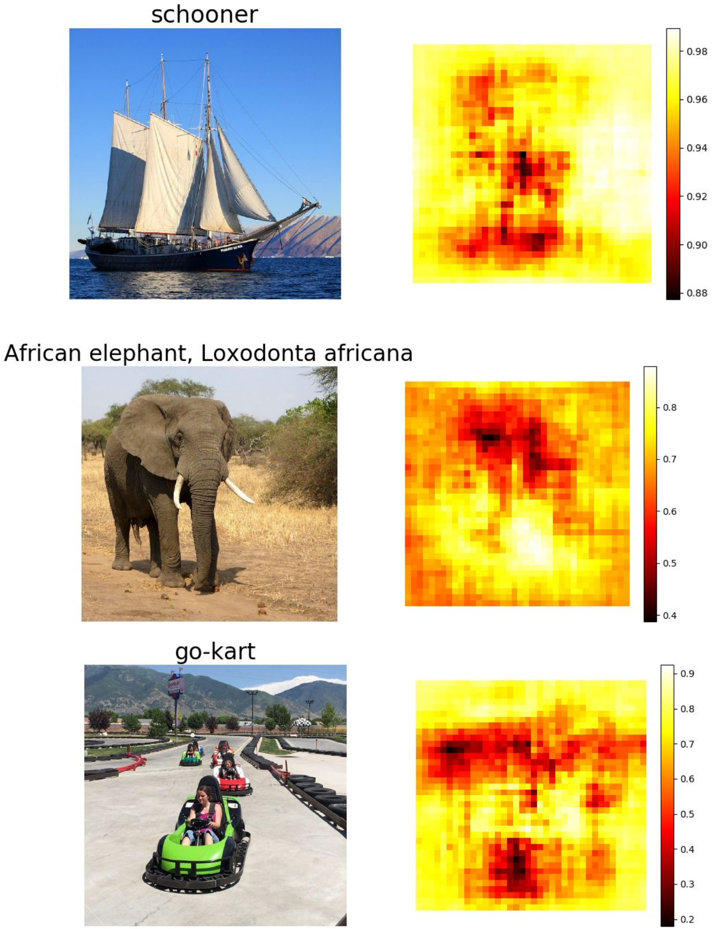




Bucinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020

#7: Unfaithful explanations: Saliency maps

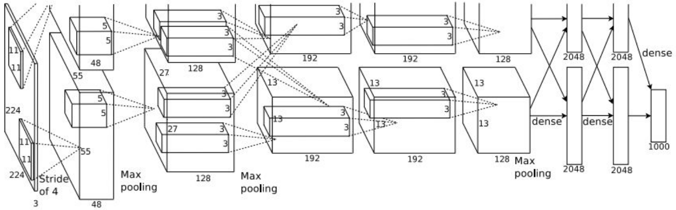




Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

Which pixels explain the prediction? Saliency via backprop

Forward pass: Compute probabilities

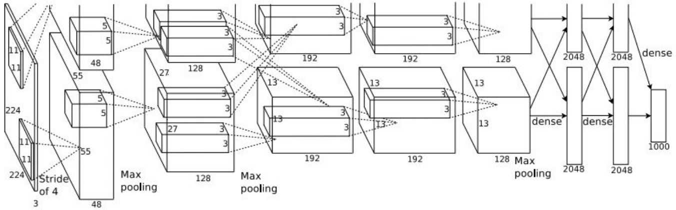


Dog

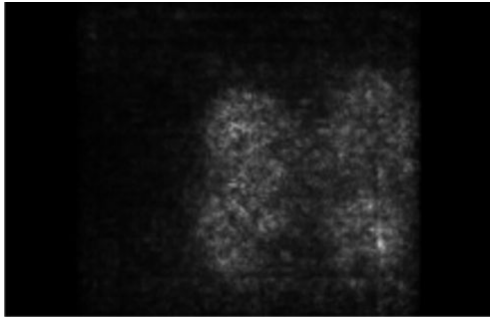
Simonyan et al. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014

Which pixels explain the prediction? Saliency via backprop

Forward pass: Compute probabilities



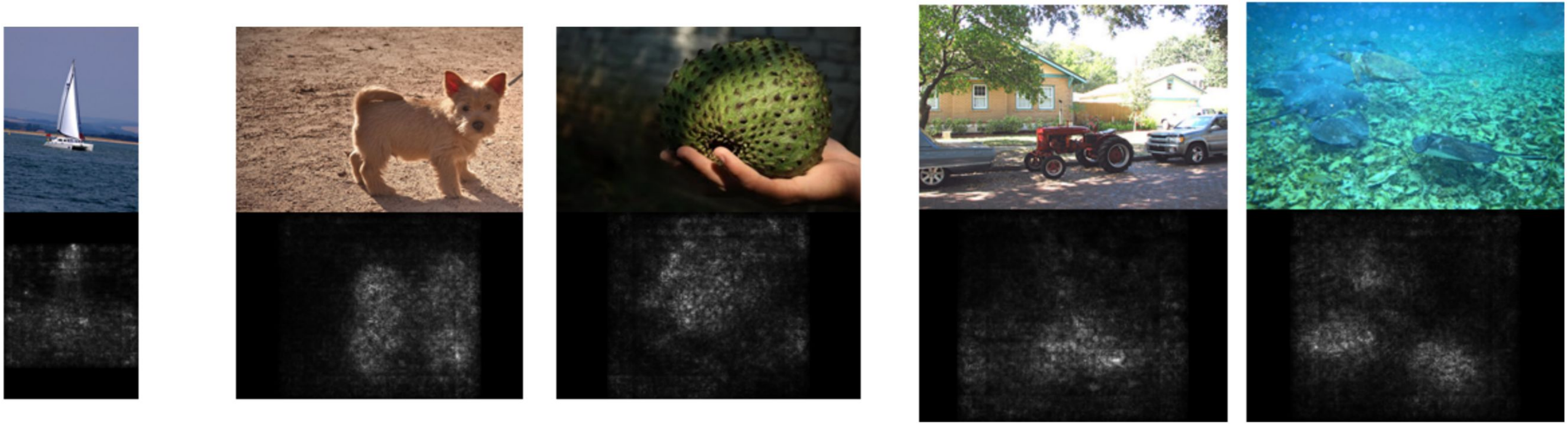
Dog



Compute gradient of (unnormalized) class score with respect to image pixels, take absolute value and max over RGB channels

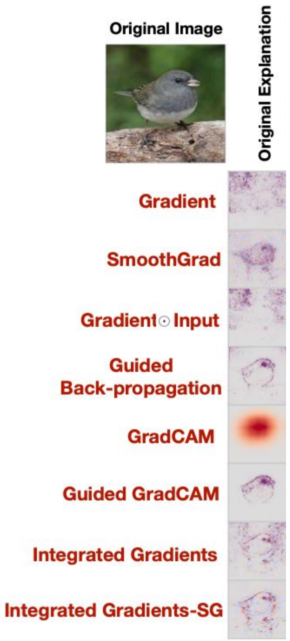
Simonyan et al. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014

Which pixels explain the prediction? Saliency via backprop



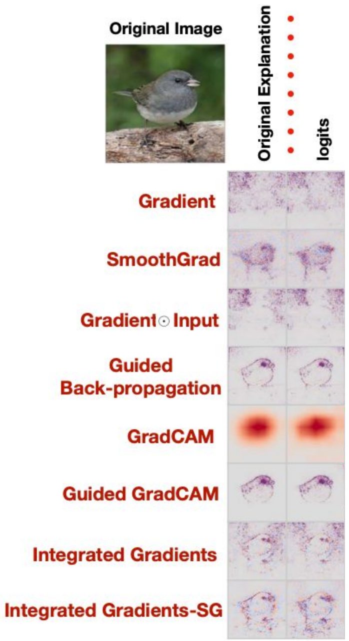
Simonyan et al. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014

Saliency maps were getting quite popular



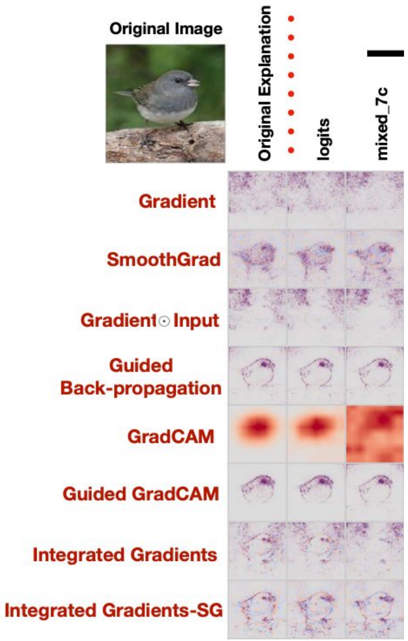
Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018

#7: Unfaithful explanations: random predictions don't change explanations



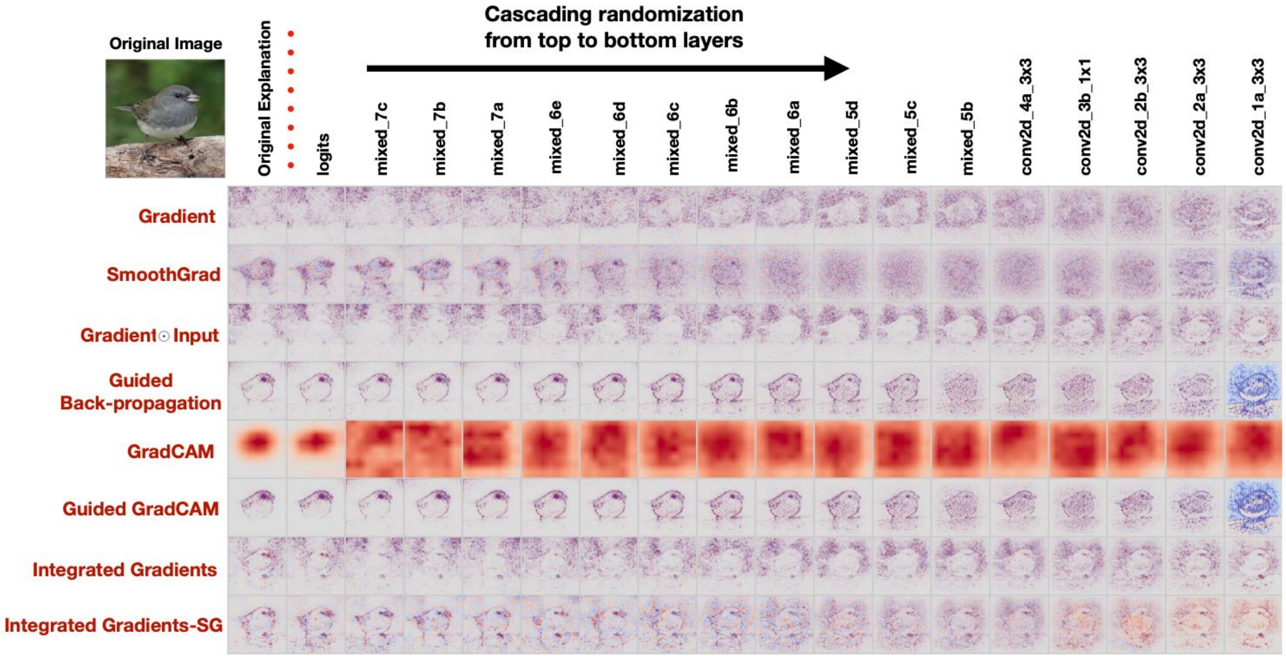
Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018

#7: Unfaithful explanations: randomizing last two layers don't change explanations



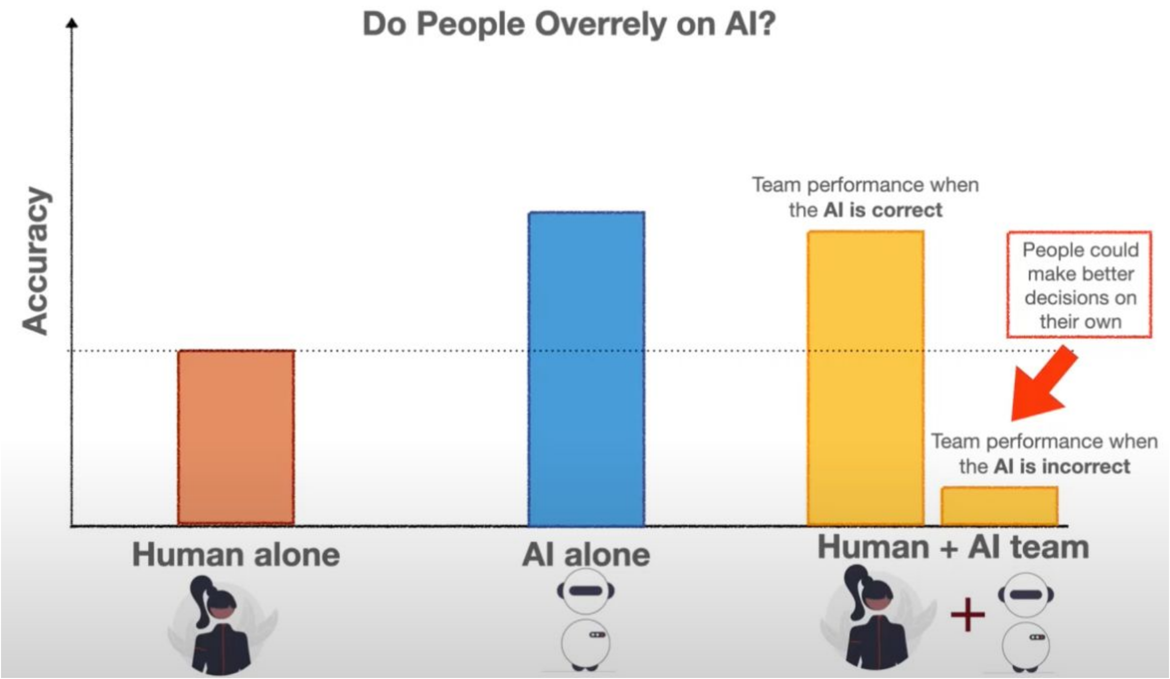
Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018

#7: Unfaithful explanations: random networks induce the same explanations



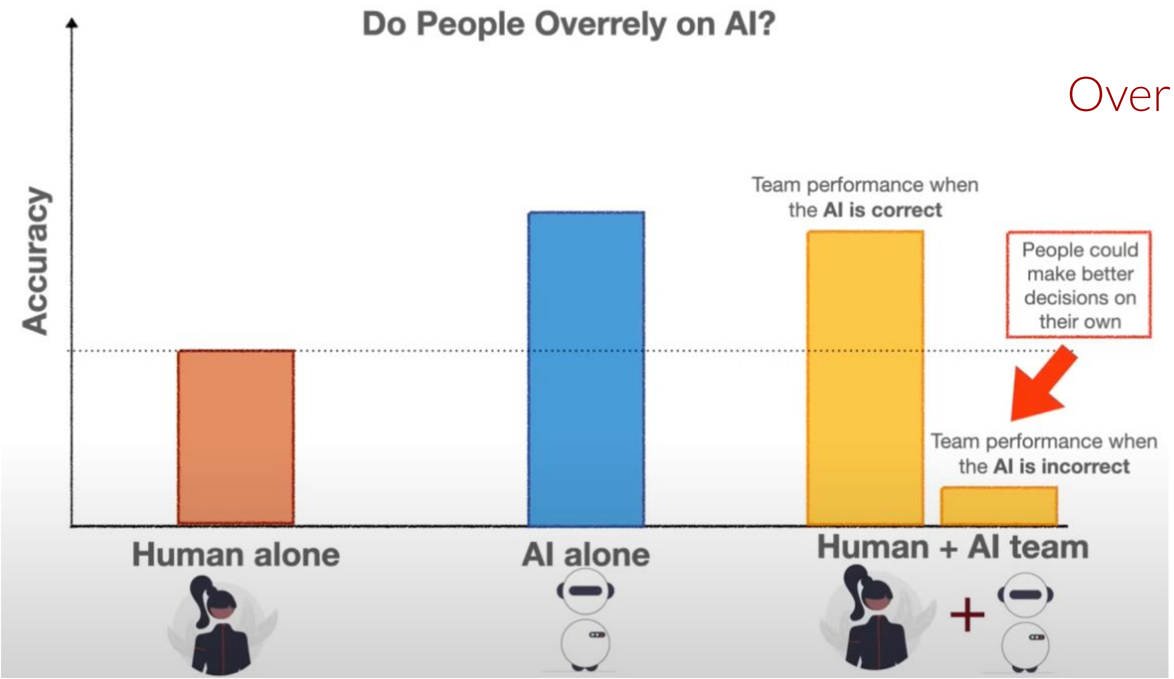
Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018

#8: Faithful explanations may still hurt decision making



Bucinca et al. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. CSCW 2021

#8: Faithful explanations may still hurt decision making



Overreliance!!!

Bucinca et al. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. CSCW 2021

Ranjay Krishna | ranjay@cs.washington.edu

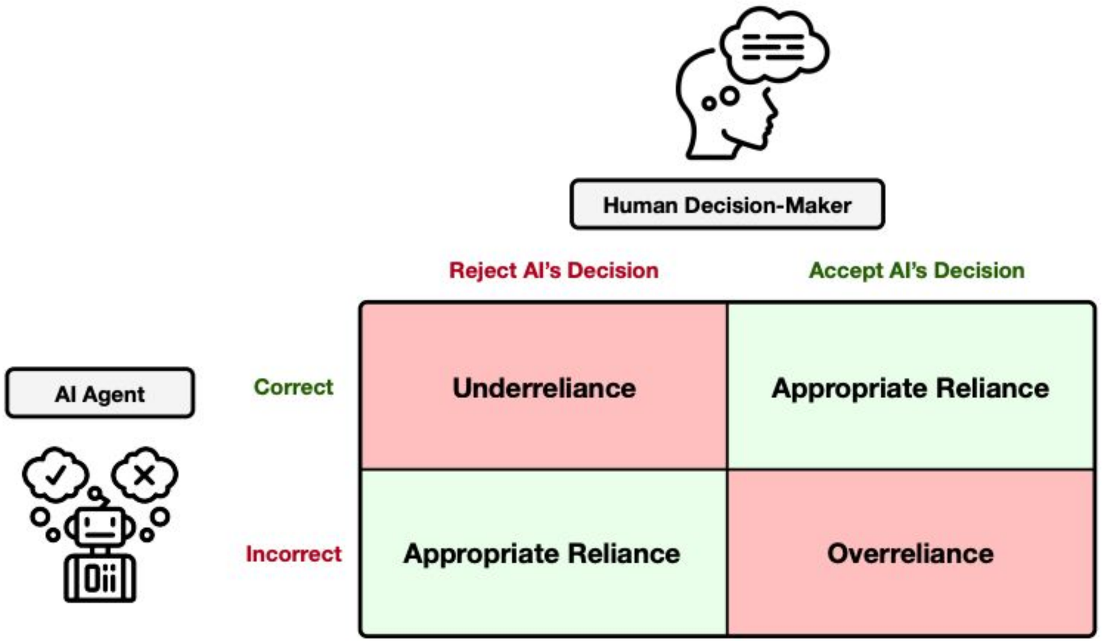
Deep Dive:

Research question:

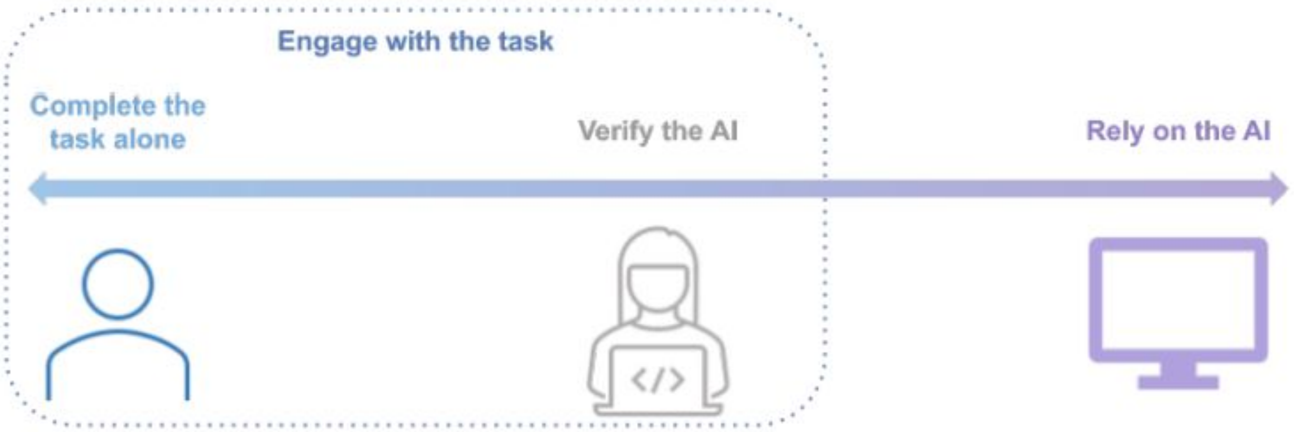
Can explanations reduce overreliance on AI-assisted decision making?

Vasconcelos et al. Explanations can reduce overreliance Overreliance on AI Systems During Decision-Making. CSCW 2023

What is overreliance?



Two prototype strategies in which people engage with explanations



Predominant hypothesis for overreliance

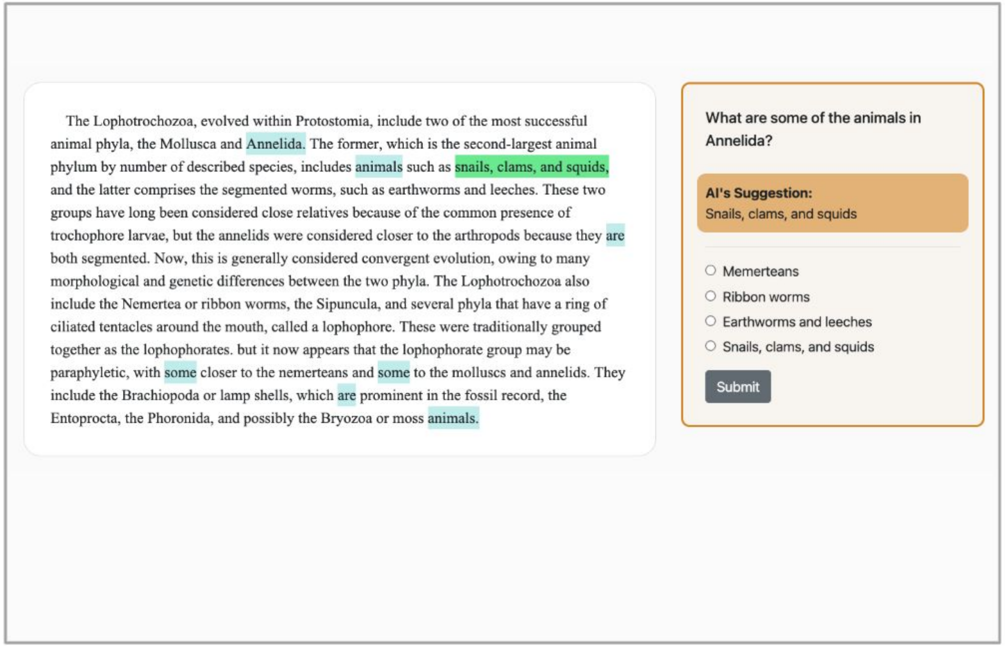
Cognitive biases

- Mere presence of explanations increase trust.
- Trust makes us overrely.

There are cases when we do engage with explanations

- Incorrect email auto-replies
- GPS navigation system showing you the wrong route
- What else have you encountered?

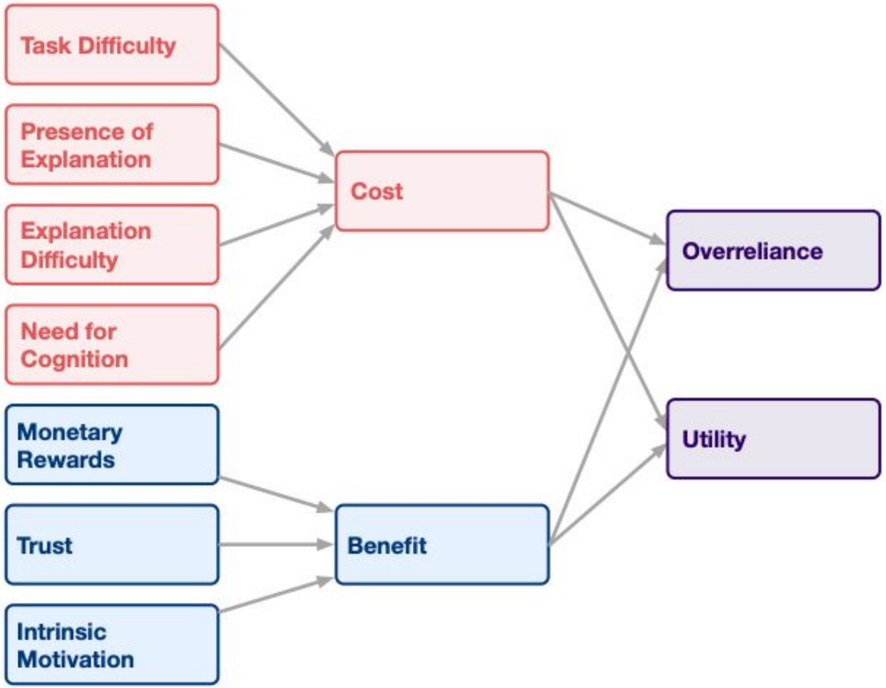
Why don’t explanations help in these tasks?



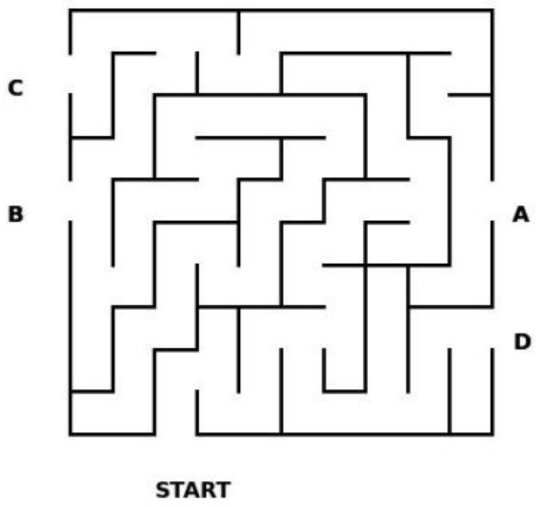
A cost-benefit framework

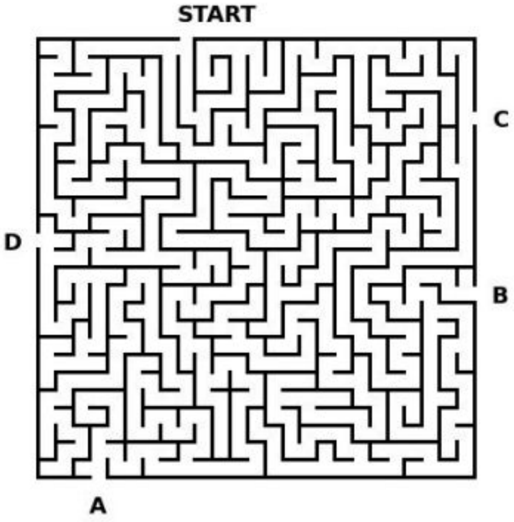
Costs increase overreliance

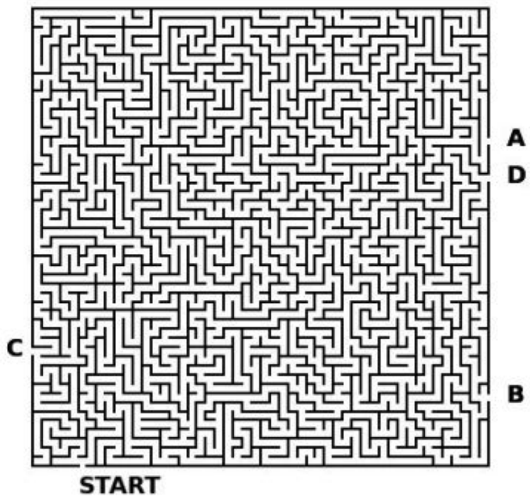
Benefits decrease



Designed tasks that increase in cognitive effort

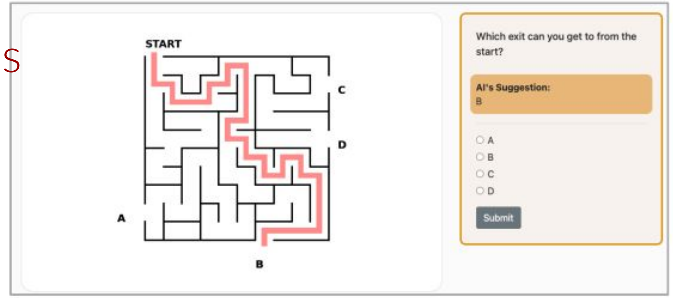




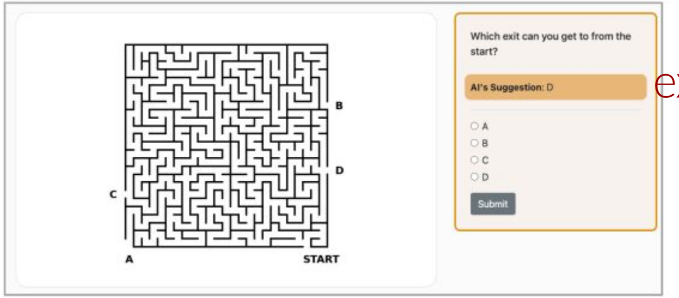


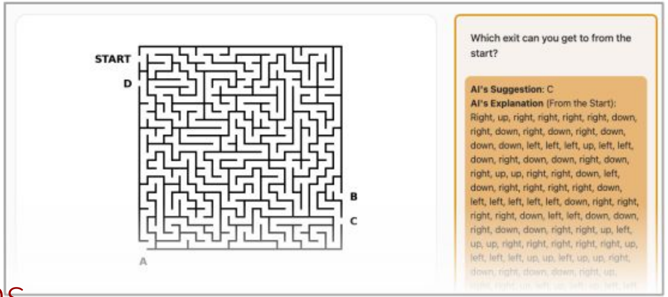
Explanations that take different cognitive effort

Highlight explanations

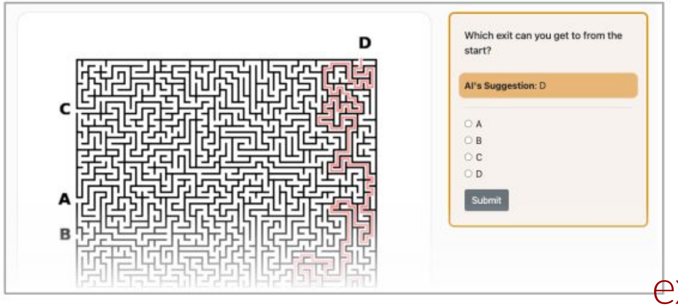


No explanations



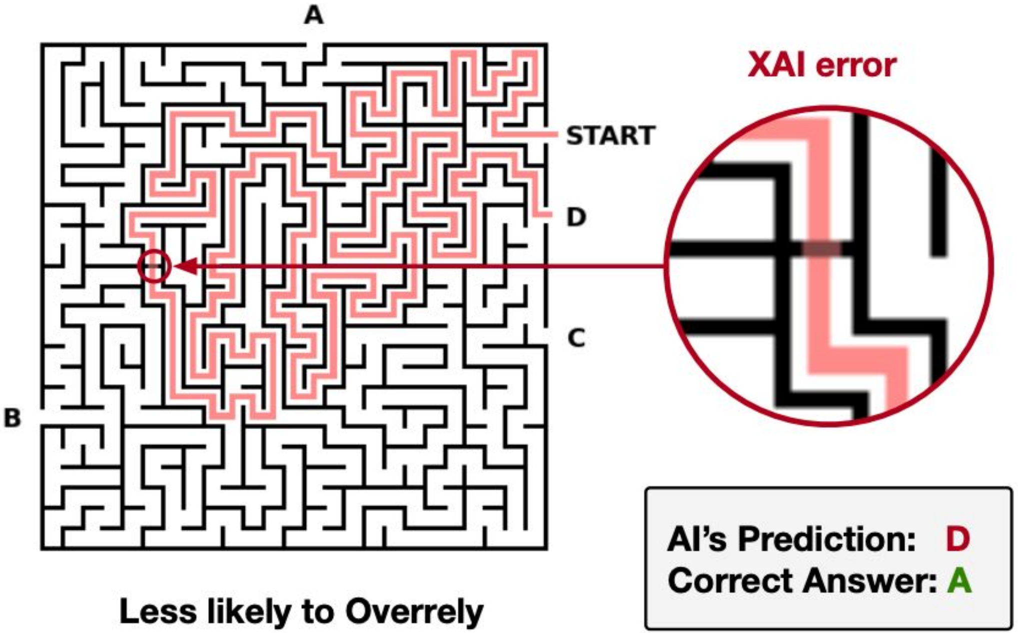


Written explanations



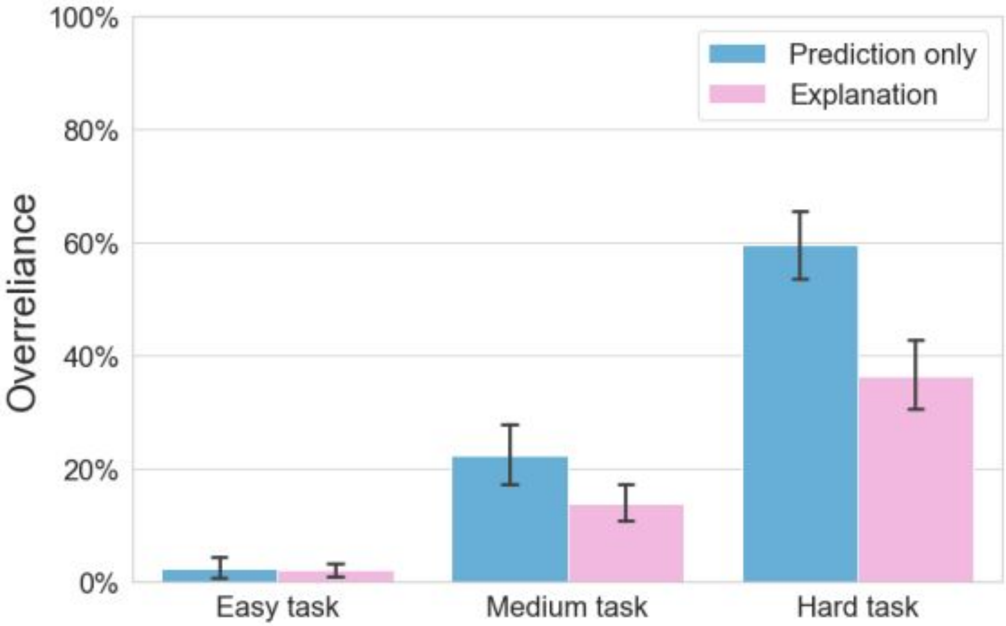
Highlight explanations for hard tasks

Highlights reduce cognitive effort to find AI errors

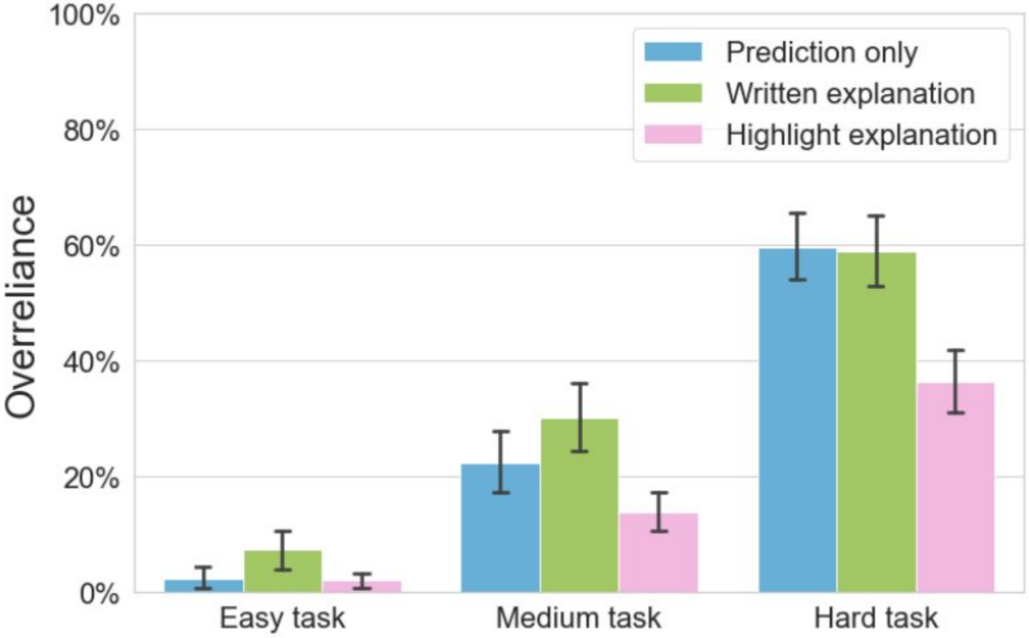


Less likely to Overrely

We show for the first time that explanations do reduce overreliance in human-AI decision making but only when the task difficulty is high enough to require explanations



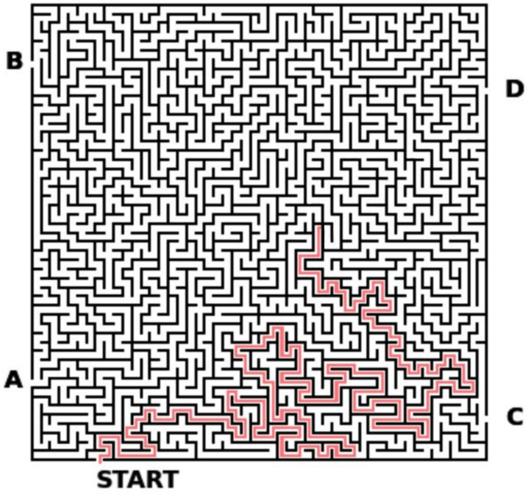
If explanations take effort to understand, overreliance increases



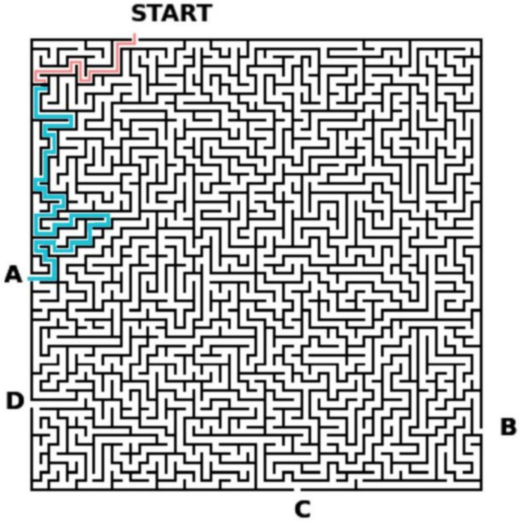
Ranjay Krishna | ranjay@cs.washington.edu

Adding two a new type of explanations:

Incomplete explanations



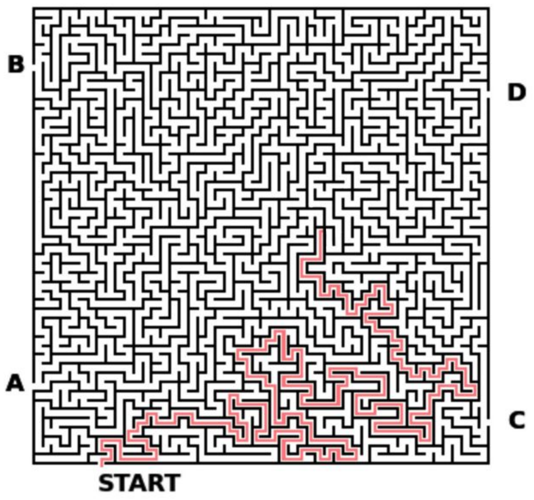
Salient explanations



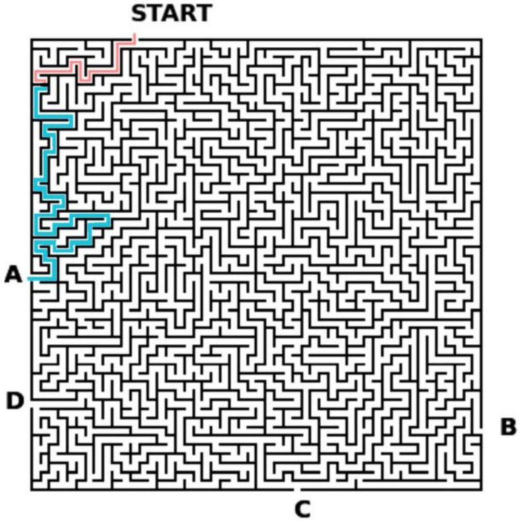
Ranjay Krishna | ranjay@cs.washington.edu

All four types of explanations? Which one do you think will have highest and lowest overreliance for hard tasks?

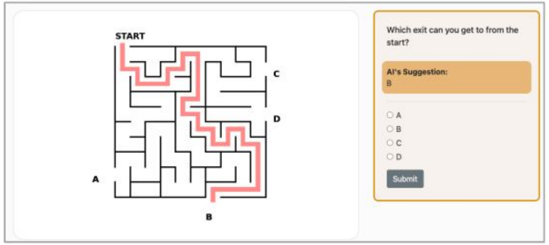
Incomplete explanations

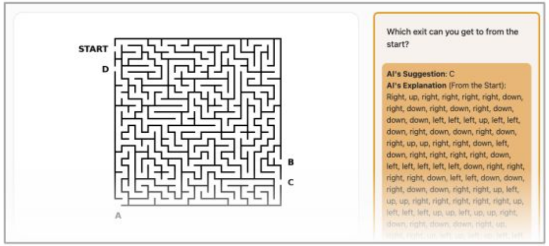


Salient explanations



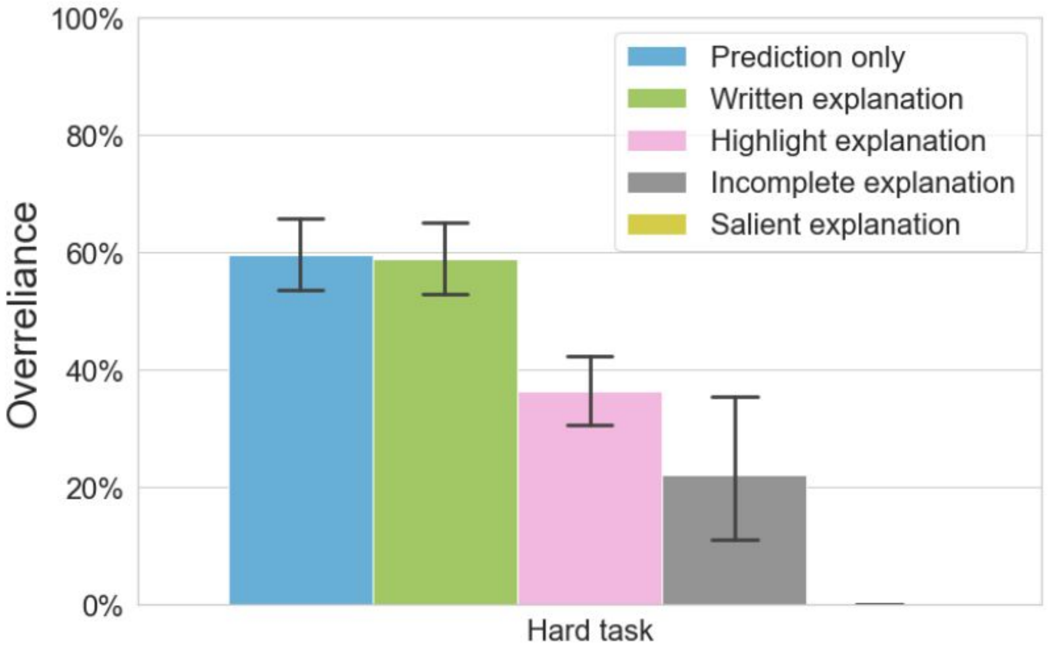
Highlight explanations



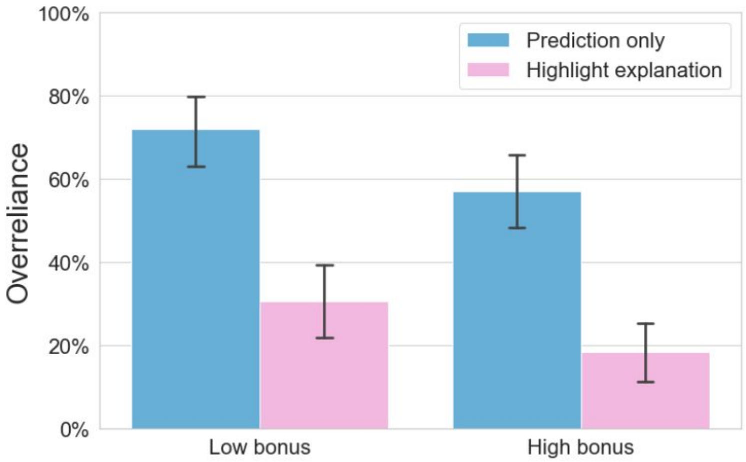


Written explanations

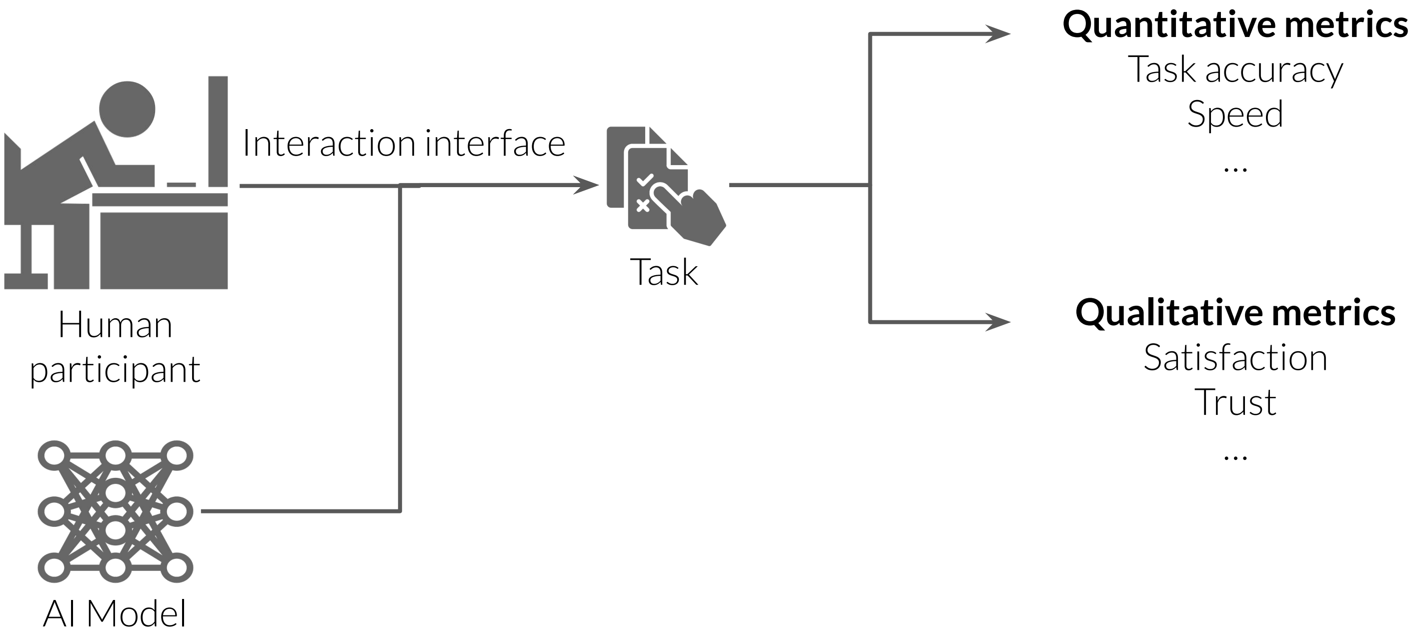
Less cognitive effort -> less overreliance



More benefit -> less overreliance



Challenges with evaluation protocols for human-AI systems



Next time:
Learning from interactions