Lecture 4 - finishing up from last time

The challenges with understanding models

Ranjay Krishna | ranjay@cs.washington.edu

Reframing with new metaphors

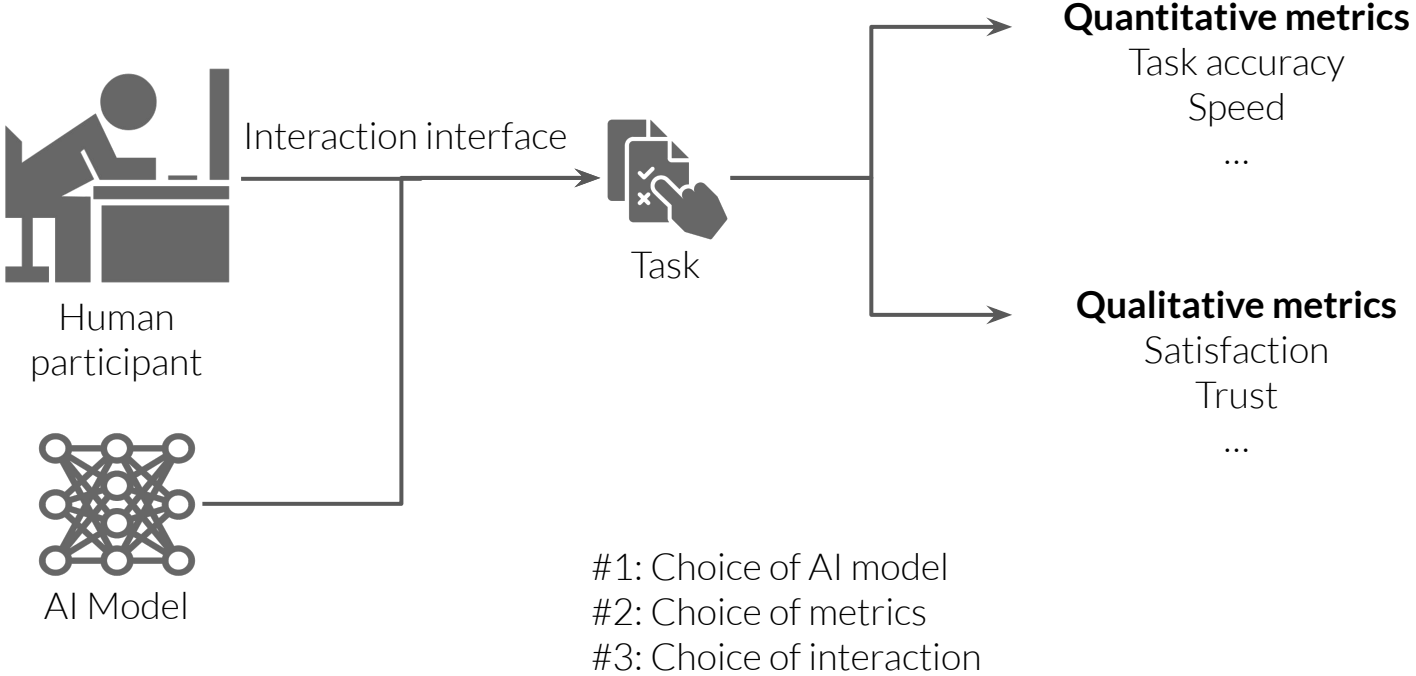|		
Intelligent Agents Manifests cognitive, linguistic, perceptual abilities	→	**Supertools** Augments human abilities and performance
Teammates Acts as a collaborator, interacts using language	→	**Tele-bots** Boosts human perception & motor skills
Assured autonomy Sets goals, makes decisions, improves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

Evaluation protocol for human-AI systems

Interaction interface

Quantitative metrics
Task accuracy
Speed
...

Task

Qualitative metrics
Satisfaction
Trust
...

Human
participant

AI Model

#1: Choice of AI model
#2: Choice of metrics
#3: Choice of interaction

What conceptual model does this language interaction afford?

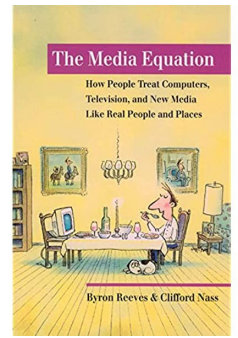Seattle space needle with neon signage in the style of bladerunner

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

Ranjay Krishna | ranjay@cs.washington.edu

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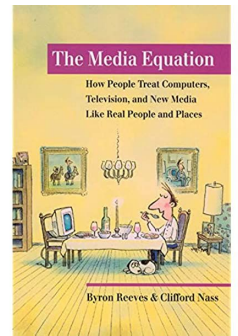
More nuanced understanding of the media equation: when machines project social competence or enable social interactions, they induce shortcut social scripts in people

- 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, ….*

Ranjay Krishna | ranjay@cs.washington.edu

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

Ranjay Krishna | ranjay@cs.washington.edu

#4: Choice of interface: The effects of anthropomorphisation

TayTweets
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

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

WARNING!!
MITSUKU WILL NOT SPEAK TO YOU IF YOU ARE ABUSIVE

THIS IS YOUR SECOND WARNING. AFTER FIVE WARNINGS YOU WILL BE BANNED FROM SPEAKING TO HER.
CLICK HERE TO CONTINUE

Ruuh's Conversations This Month

Total number of messages received
1,239,446

Total number of insults and abuses received
94,392

This Chatbot has Over 660 Million Users—and It Wants to Be Their Best Friend

#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?

Ranjay Krishna | ranjay@cs.washington.edu

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

AMT workers → Consent and study instructions → Metaphor displayed (Toddler) → Pre-use expectation measurement → Interaction with system → User evaluation and gauging user attitude

Study Workflow

AMT workers → Consent and study instructions → Metaphor displayed → Pre-use expectation measurement → Interaction with system → User evaluation and gauging user attitude
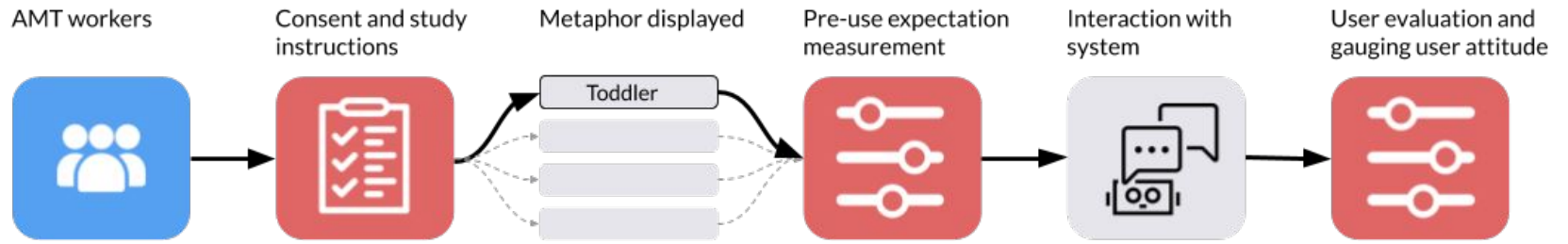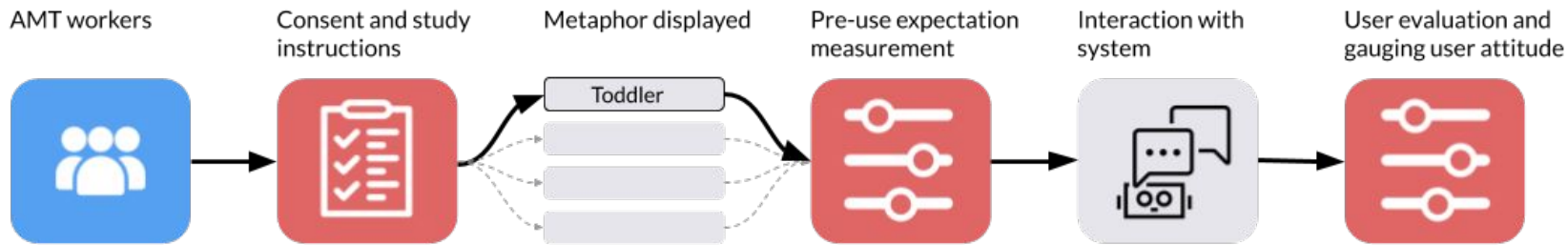
How do you choose the metaphors?

High competence

Low warmth

Low warmth
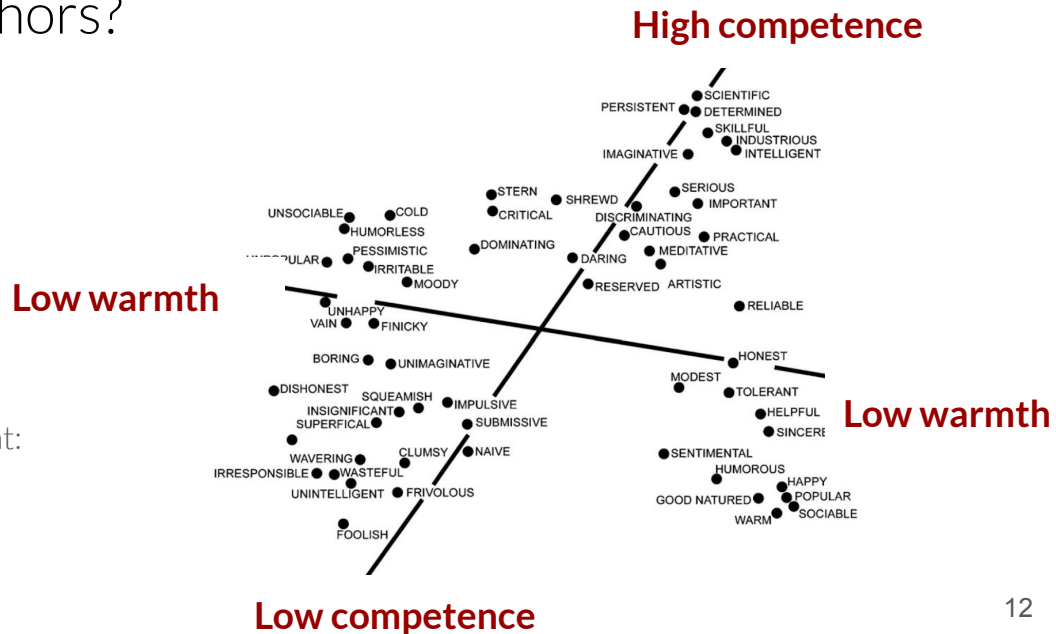
Low competence

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

Hi

Hello, I am planning for a trip to New York from Montreal

We have options for you.

Luma Hotel - 4 Star: Dates Available: August 26 to 30 (5 nights) Price: $850.90

Yes

is it available on that dates?

yeah! that sounds good.

Okay. I'll book that for you right away.

Do you want me to show you your return flight options?

yes

United - Departure at 2:00 pm local time - Direct Flight - $311

Does berlin offer parking?

I am sorry. I don't have that information.

Great. it will work.

| AI system | User |

We sampled metaphors along these two dimensions
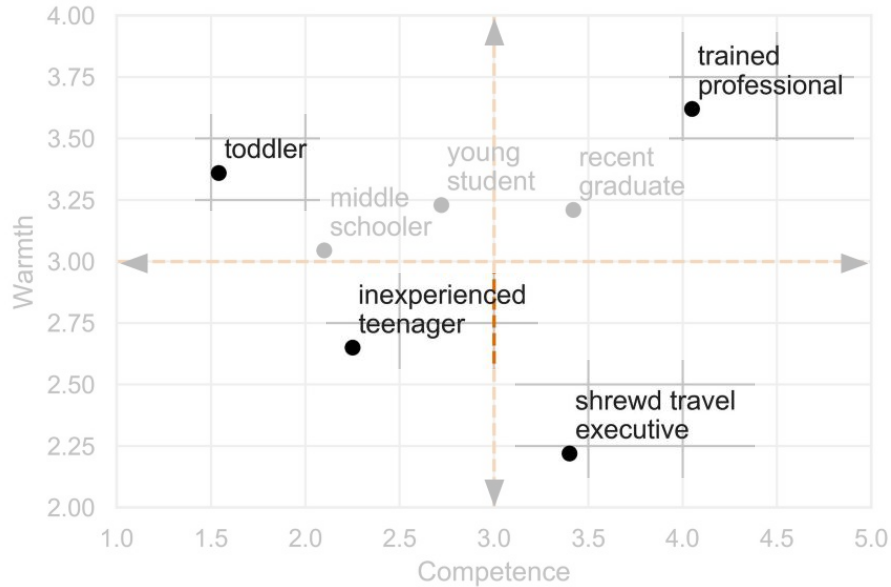
Manipulations:
4 treatment Groups + 1 Control Group that is not shown a metaphor

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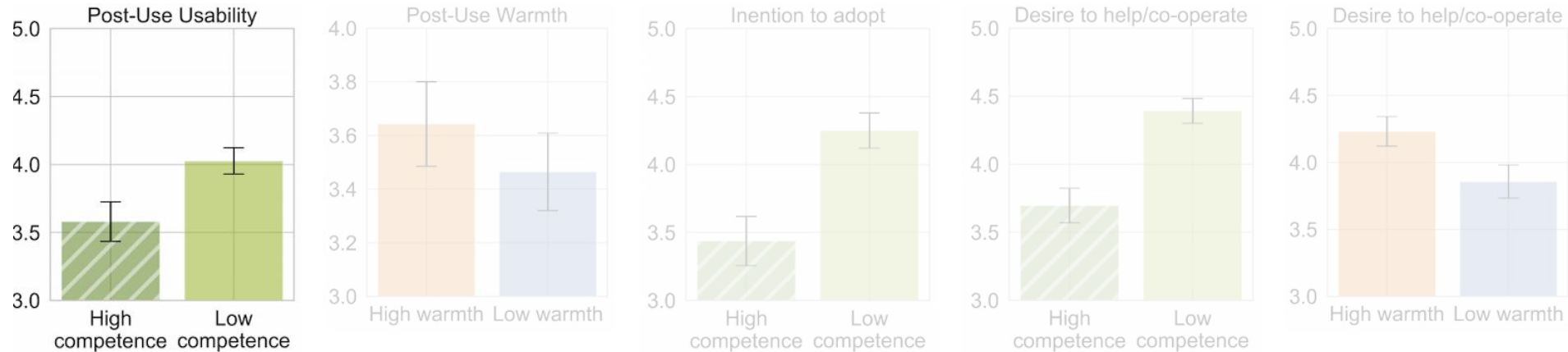
- Support for H2 and Contrast Theory - over performing expectations leads to positive evaluations

Metaphors directly affect how warm people think models are
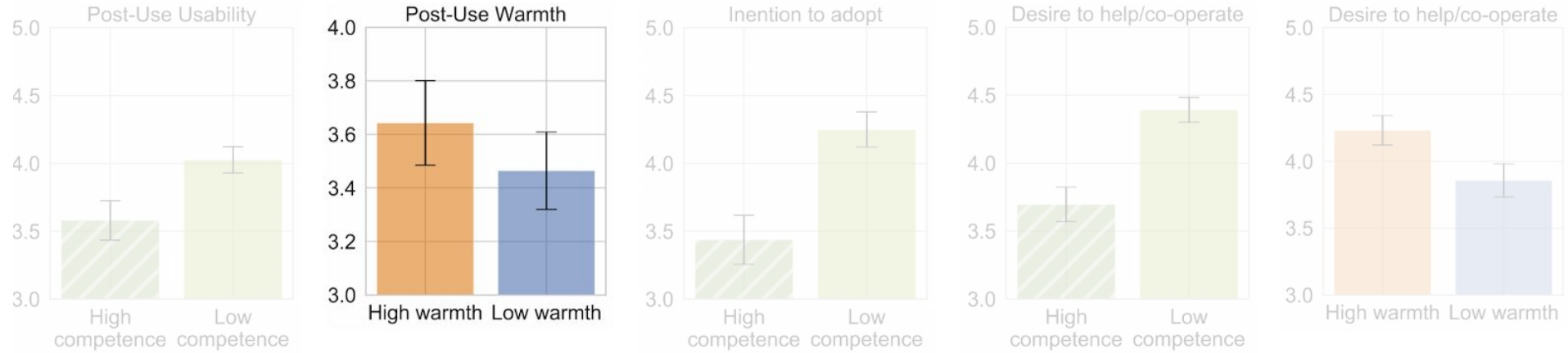
Low competence metaphors increase users' likelihood of adopting the AI agent
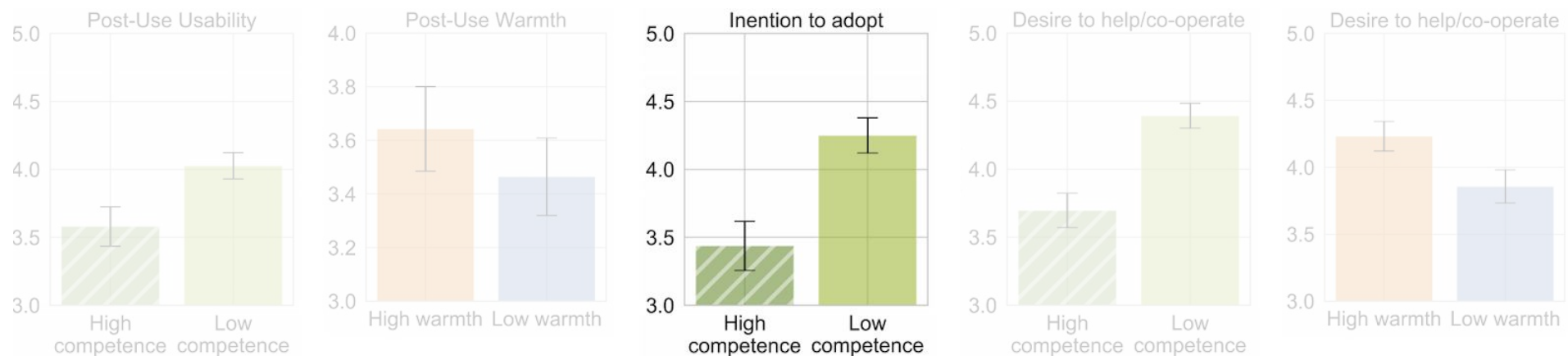
Users prefer to cooperate with agents that have high warmth and low competence
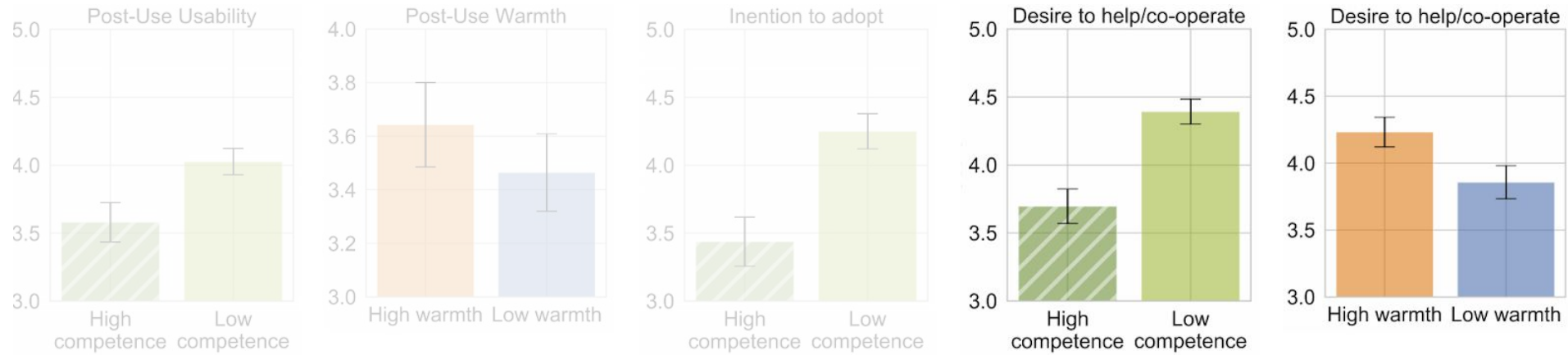
- mixed support to both H1 and H2:
  - assimilation theory along the warmth dimension
  - contrast theory along the competence dimension.

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

High warmth conversation

Is wifi included?

yes

And the Paris location does that include breakfast?

Breakfast is offered

Do they speak english?

I am sorry. I don't have that information.

Do I get my own bathroom?

I am sorry. I don't have that information.

How far from the Empire State Building is the New Yorkhotel?

I am sorry. I don't have that information.

Does berlin offer parking?

I am sorry. I don't have that information.

Do any of these hotels offer spa services?

yes

New York and Paris

Do any have a minibar?

I am sorry. I don't have that information.

I would like to book the Paris hotel.

Low warmth conversation

Hi

Hello, I am planning for a trip to New York from Montreal

We have options for you.

Luma Hotel - 4 Star: Dates Available: August 26 to 30 (5 nights) Price: $850.90

Yes

is it available on that dates?

yeah! that sounds good.

Okay. I'll book that for you right away.

Do you want me to show you your return flight options?

yes

United - Departure at 2:00 pm local time - Direct Flight - $311

Does berlin offer parking?

I am sorry. I don't have that information.

Great. it will work.

| AI system | User |

effect is greater as the violation is greater

Adoption intention v/s Competence
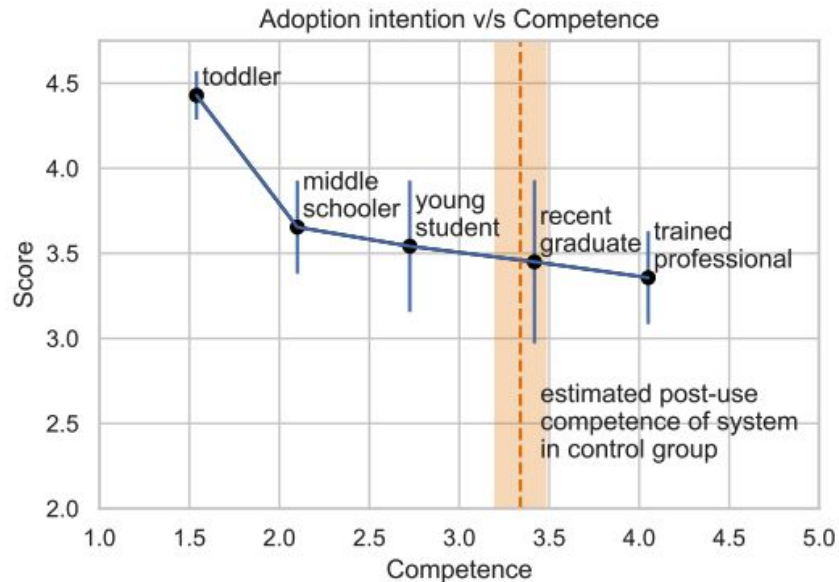
Extreme violations of expectations have stronger effects

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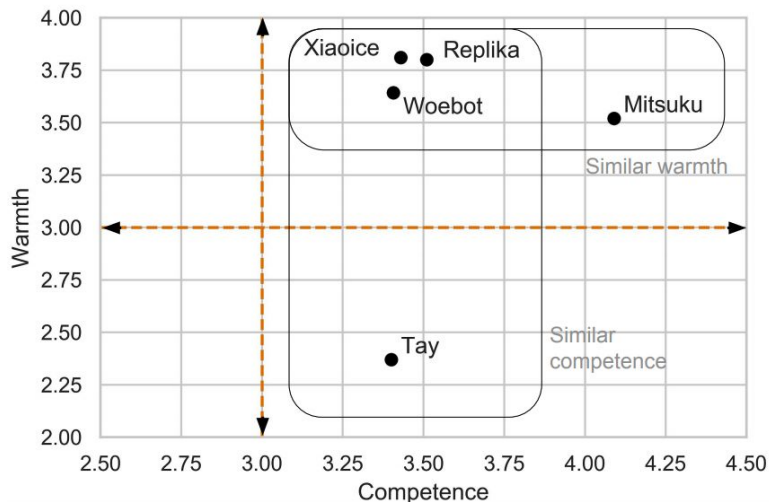
Most chabots today signal high competence. => users are left disappointed

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.

Website User Survey

1. The website has a user friendly interface.

strongly agree — agree — neutral — disagree — strongly disagree
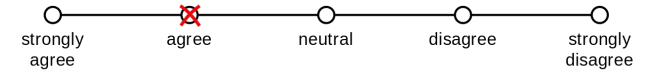
2. The website is easy to navigate.

strongly agree — agree — neutral — disagree — strongly disagree
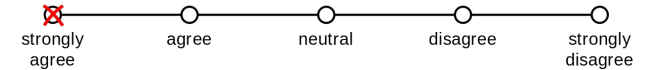
3. The website's pages generally have good images.

strongly agree — agree — neutral — disagree — strongly disagree
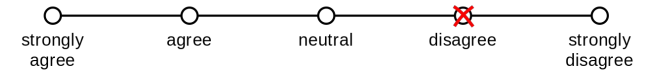
4. The website allows users to upload pictures easily.

strongly agree — agree — neutral — disagree — strongly disagree
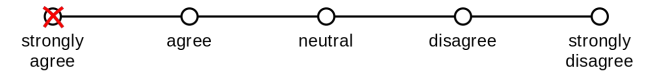
5. The website has a pleasing color scheme.
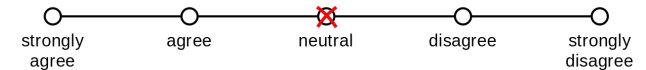
strongly agree — agree — neutral — disagree — strongly disagree

Ranjay Krishna | ranjay@cs.washington.edu

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

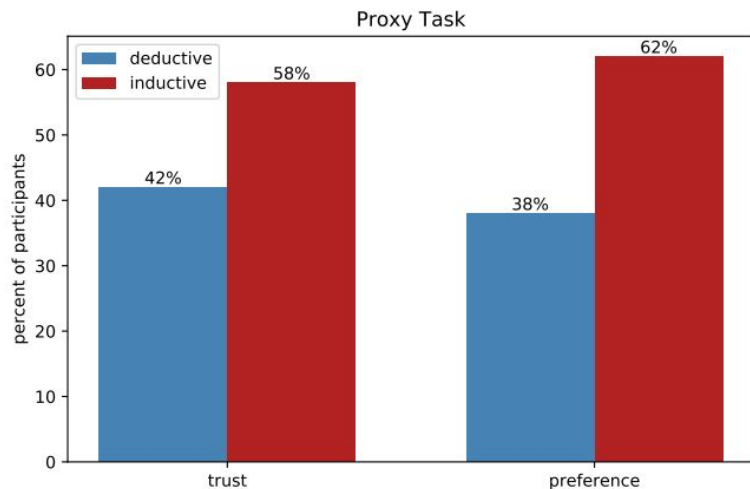Is 30% or more of the nutrients on this plate fat?

NO, 30% or more of the nutrients on this plate is not 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

Ranjay Krishna | ranjay@cs.washington.edu

#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.

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

Here are examples of plates that the AI categorizes as similar to the one above and do not have 30% or more fat:

This AI recommended answer is:

NO, 30% or more of the nutrients on this plate is not 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

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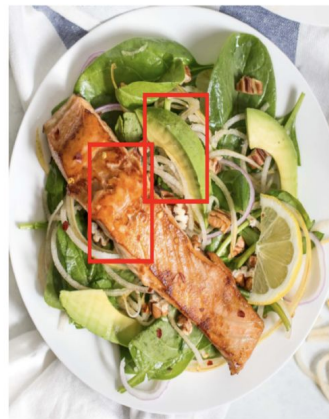
Here are ingredients the AI recognized as main nutrients which make up 30% or more fat on this plate:

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

Ranjay Krishna | ranjay@cs.washington.edu

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

Proxy 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

Ranjay Krishna | ranjay@cs.washington.edu

#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
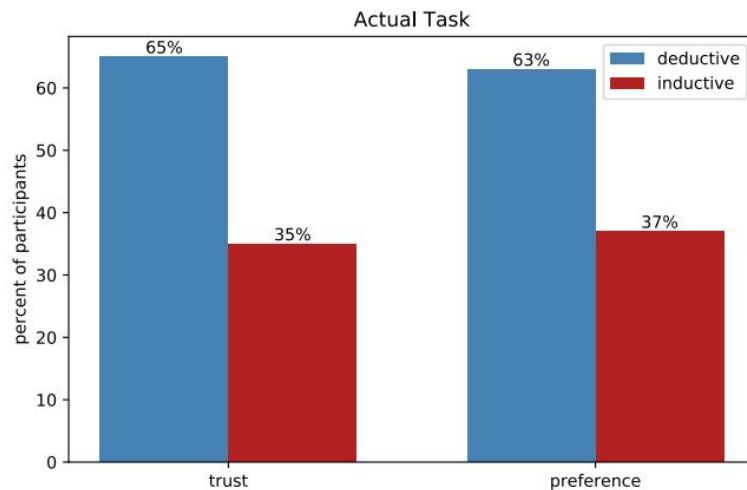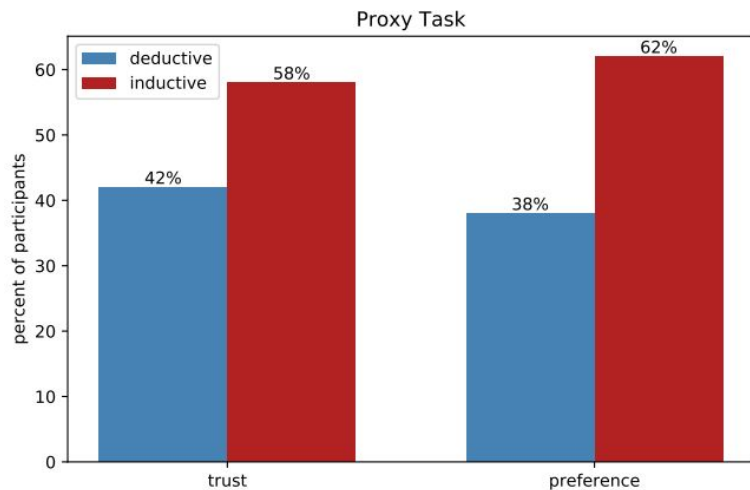
#7: Unfaithful explanations:
Saliency maps
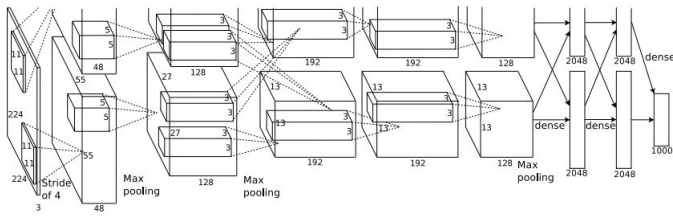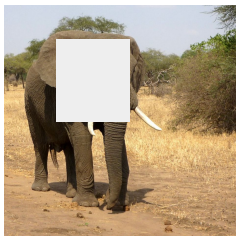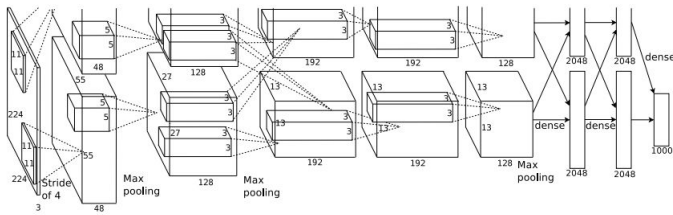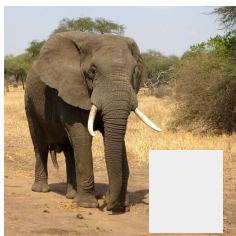
schooner
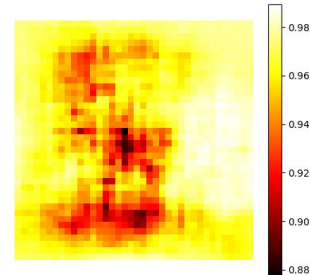
African elephant, Loxodonta africana
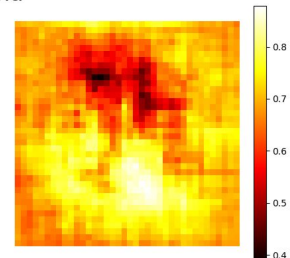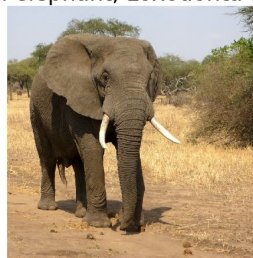
go-kart
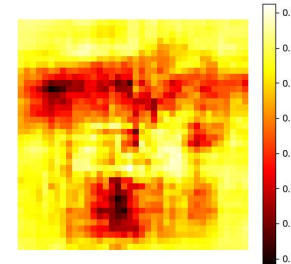
Ranjay Krishna | ranjay@cs.washington.edu

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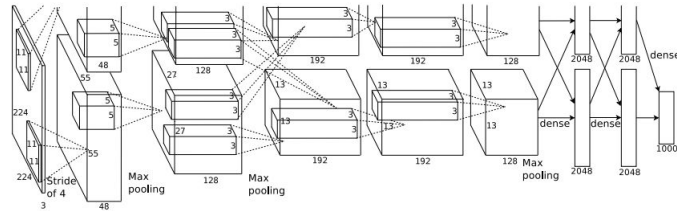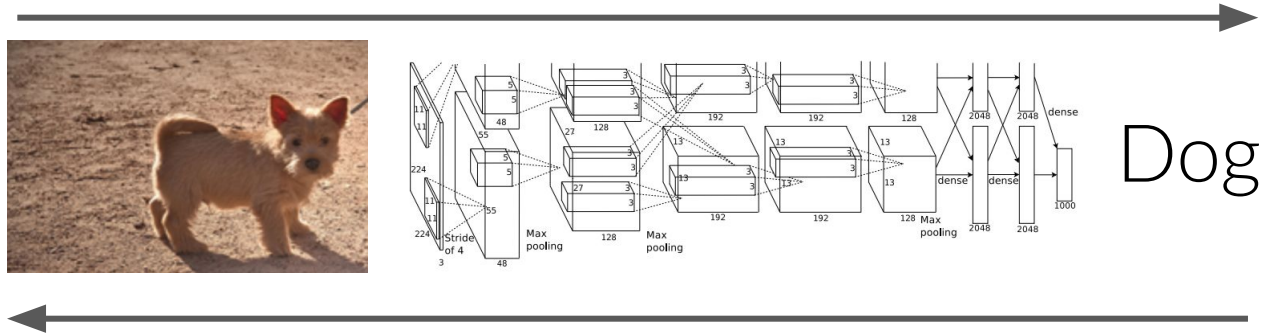
Compute gradient of (unnormalized) class score with respect to image pixels, take absolute value and max over RGB channels

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

Ranjay Krishna | ranjay@cs.washington.edu

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
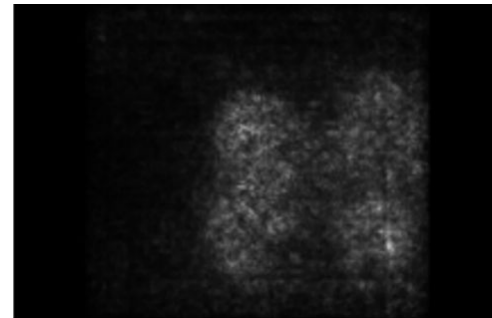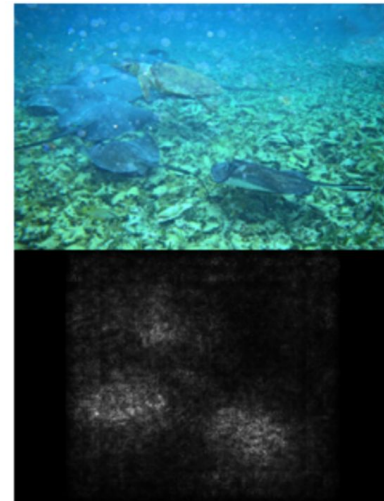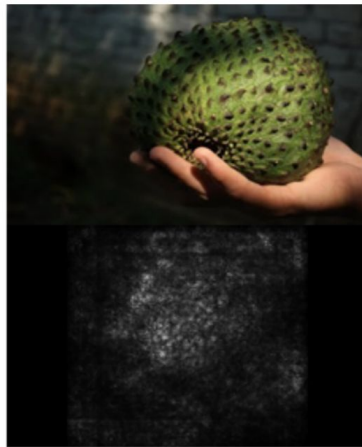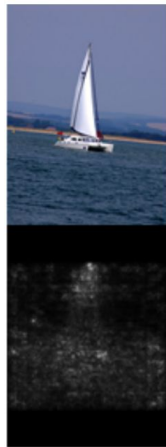
Saliency maps were getting quite popular

Original Image

Original Explanation

Gradient

SmoothGrad

Gradient⊙Input

Guided
Back-propagation

GradCAM

Guided GradCAM

Integrated Gradients

Integrated Gradients-SG

Ranjay Krishna | ranjay@cs.washington.edu

Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018
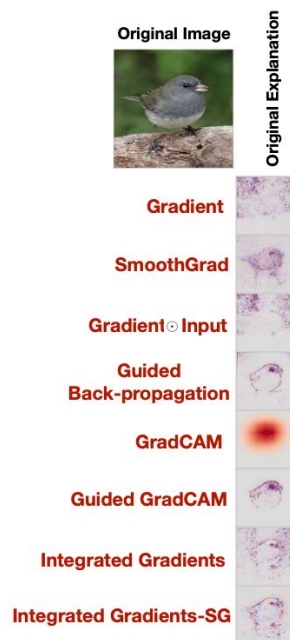
#7: Unfaithful explanations: random predictions don't change explanations

Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018
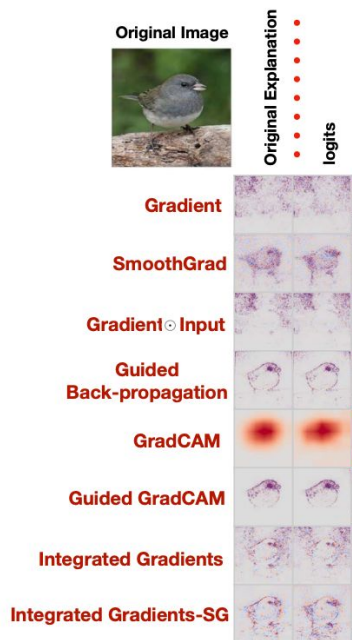
#7: Unfaithful explanations: randomizing last two layers don't change explanations

Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018
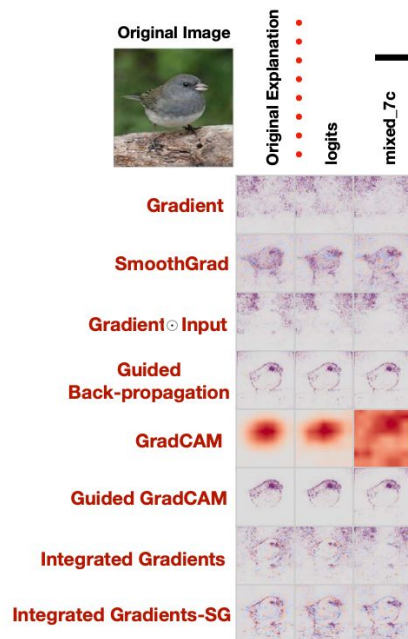
#7: Unfaithful explanations: random networks induce the same explanations
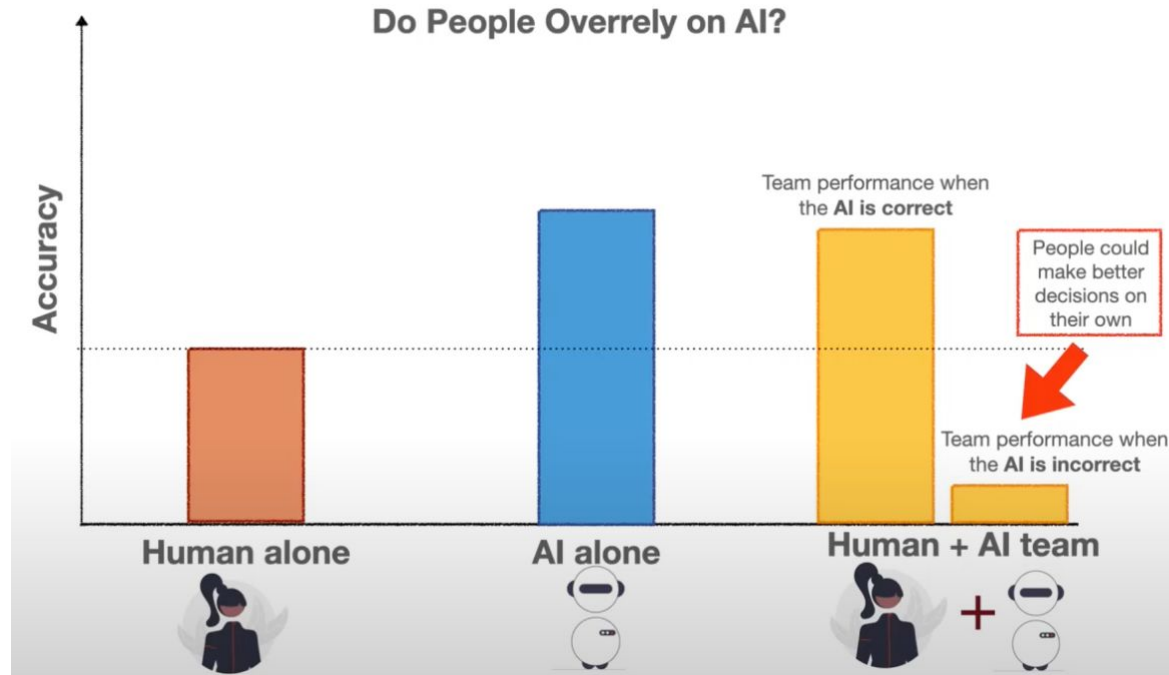
Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018

#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

Ranjay Krishna | ranjay@cs.washington.edu

#8: Faithful explanations may still hurt decision making

Overreliance!!!

Do People Overrely on AI?

Team performance when the AI is correct

People could make better decisions on their own

Team performance when the AI is incorrect

Accuracy

Human alone   AI alone   Human + AI team

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
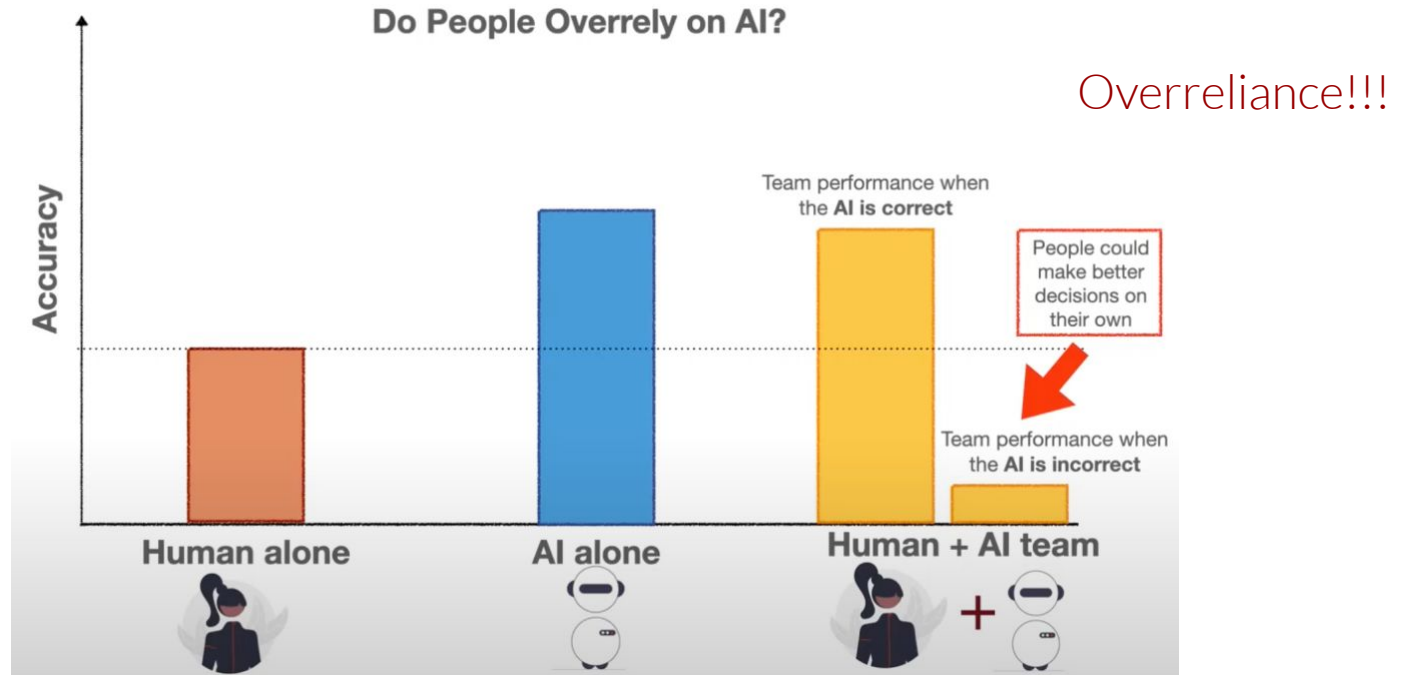
Deep Dive:

Research question:

Can explanations reduce overreliance on AI-assisted decision making?

Ranjay Krishna | ranjay@cs.washington.edu

What is overreliance?
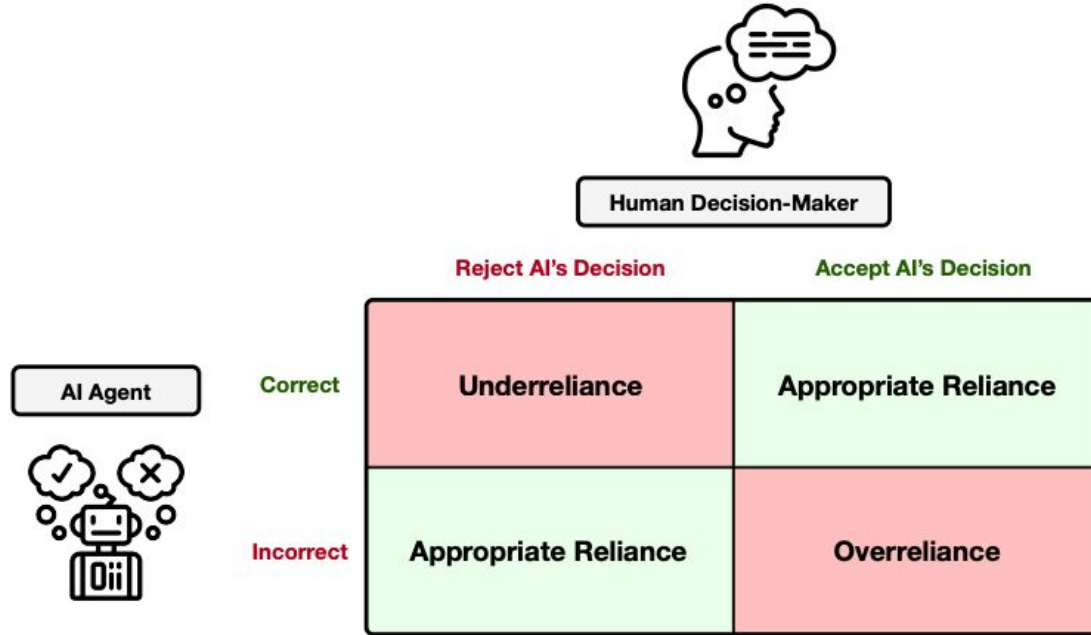
Two prototype strategies in which people engage with explanations

Engage with the task

Complete the task alone

Verify the AI

Rely on the AI

Ranjay Krishna | ranjay@cs.washington.edu
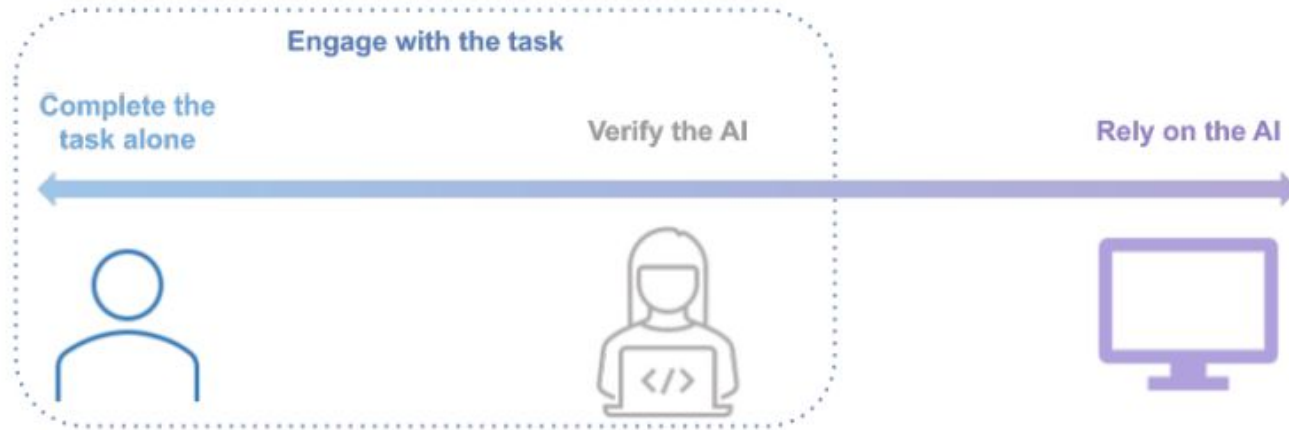
Predominant hypothesis for overreliance

Cognitive biases

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

Ranjay Krishna | ranjay@cs.washington.edu

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?

The AI must decide: Is 30% or more of the nutrients on this plate fat?

Fact: 30% or more of the nutrients on this plate is not fat.

Here are examples of plates that the AI knows the fat content of and categorizes as similar to the one above:

What will the AI decide?

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

The Lophotrochozoa, evolved within Protostomia, include two of the most successful animal phyla, the Mollusca and Annelida. The former, which is the second-largest animal phylum by number of described species, includes animals such as snails, clams, and squids, and the latter comprises the segmented worms, such as earthworms and leeches. These two groups have long been considered close relatives because of the common presence of trochophore larvae, but the annelids were considered closer to the arthropods because they are both segmented. Now, this is generally considered convergent evolution, owing to many morphological and genetic differences between the two phyla. The Lophotrochozoa also include the Nemertea or ribbon worms, the Sipuncula, and several phyla that have a ring of ciliated tentacles around the mouth, called a lophophore. These were traditionally grouped together as the lophophorates. but it now appears that the lophophorate group may be paraphyletic, with some closer to the nemerteans and some to the molluscs and annelids. They include the Brachiopoda or lamp shells, which are prominent in the fossil record, the Entoprocta, the Phoronida, and possibly the Bryozoa or moss animals.

What are some of the animals in Annelida?

AI's Suggestion:
Snails, clams, and squids

○ Memerteans
○ Ribbon worms
○ Earthworms and leeches
○ Snails, clams, and squids

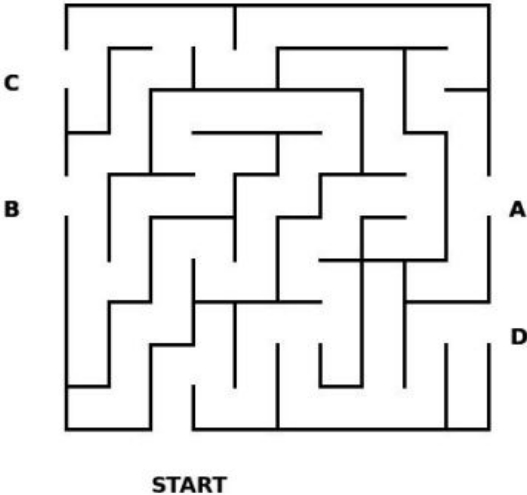Submit

Ranjay Krishna | ranjay@cs.washington.edu

A cost-benefit framework
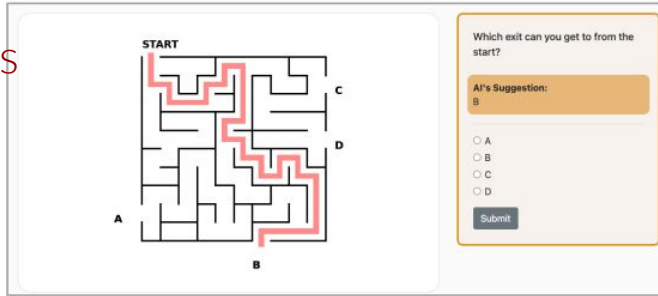
Costs increase overreliance

Benefits decrease
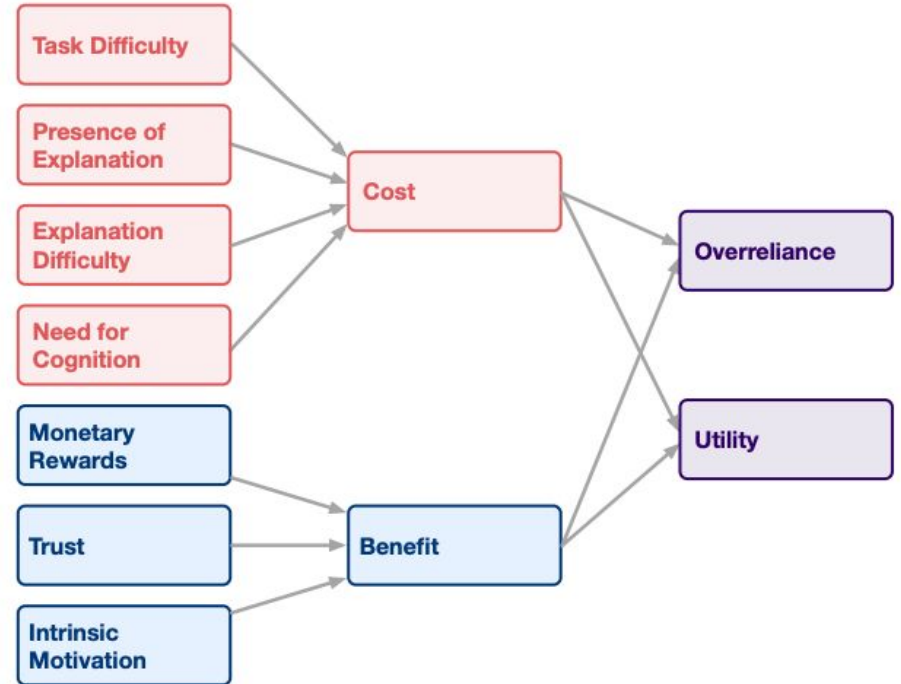
Designed tasks that increase in cognitive effort
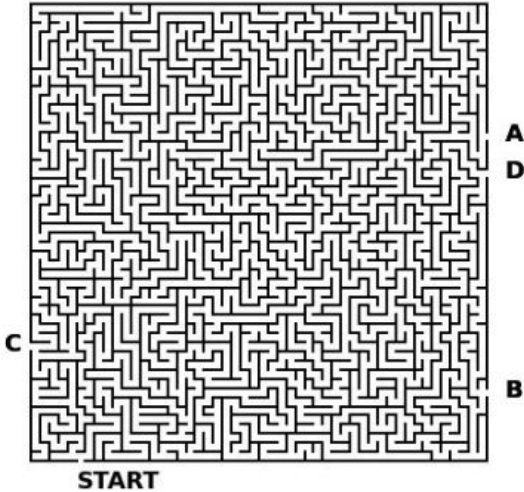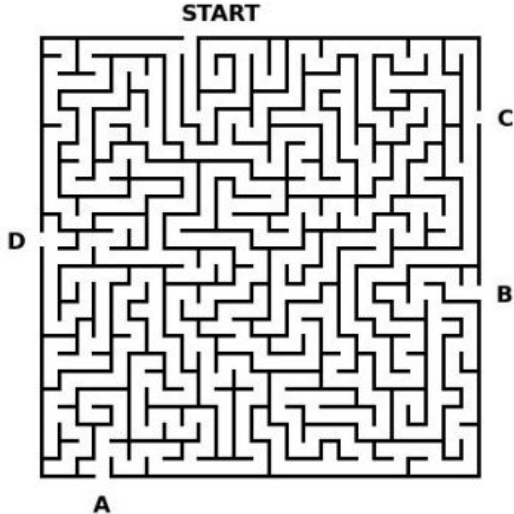
Explanations that take different cognitive effort
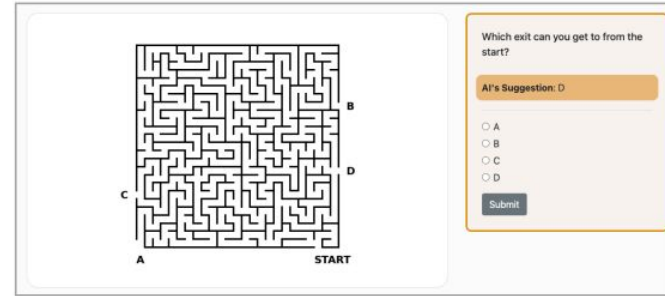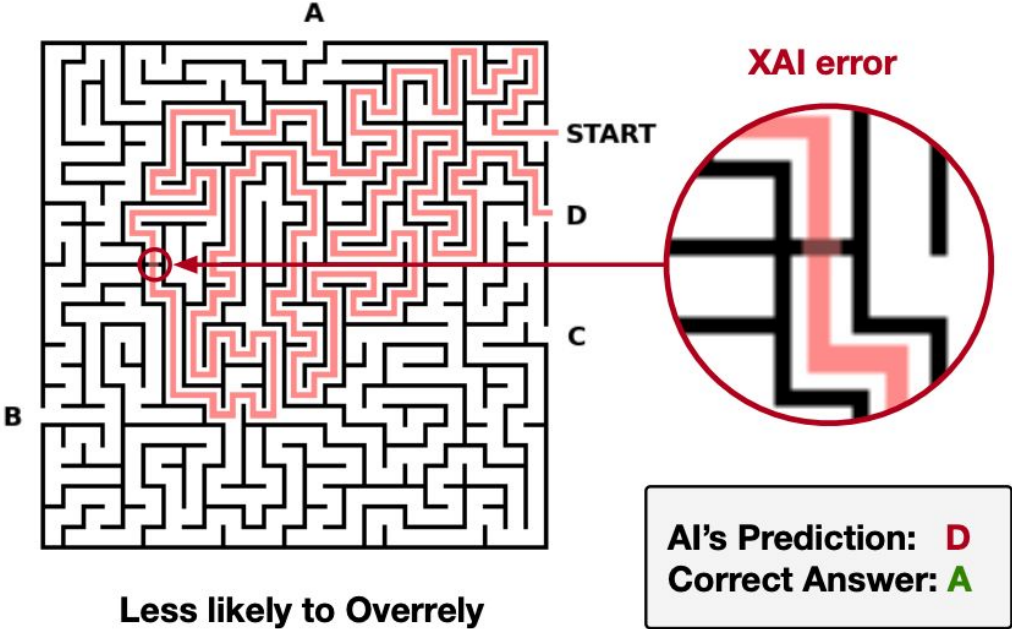
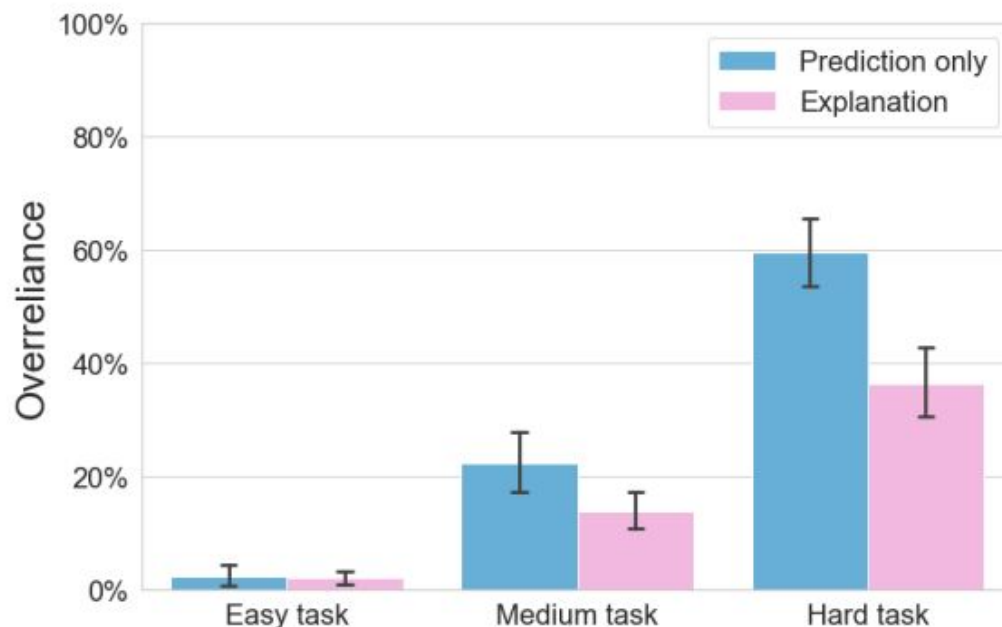Highlight explanations

No explanations

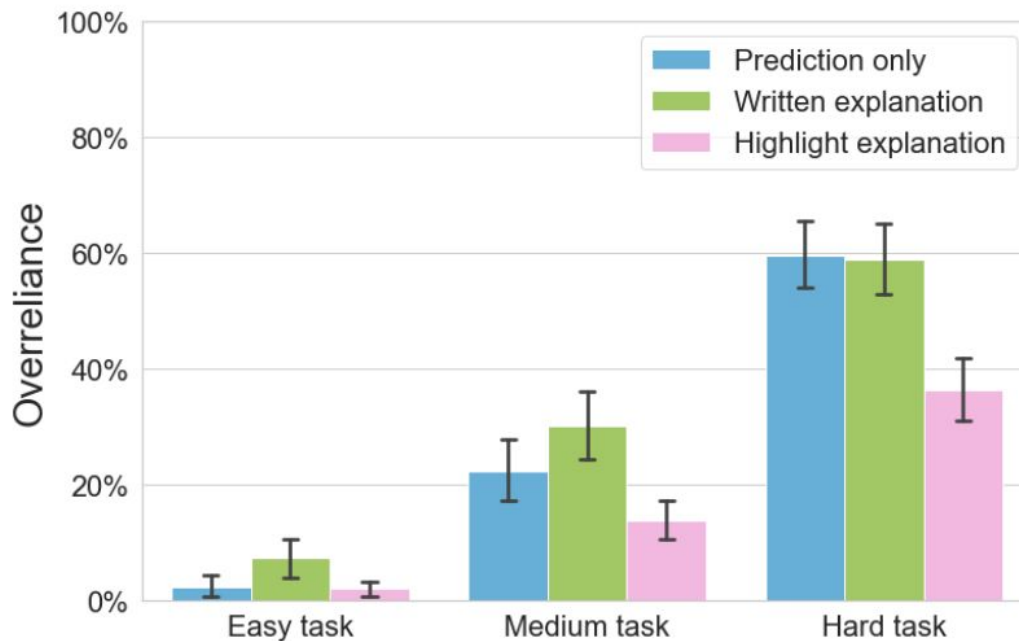Written explanations

Highlight explanations for hard tasks

Highlights reduce cognitive effort to find AI errors

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
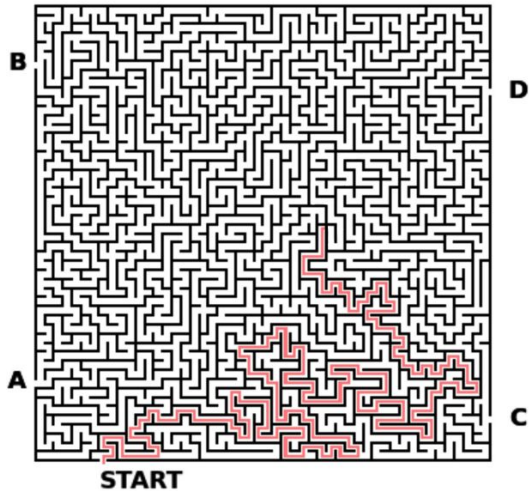
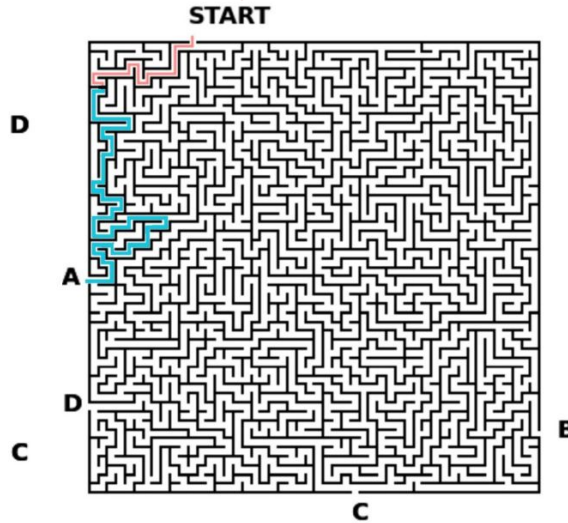# If explanations take effort to understand, overreliance increases
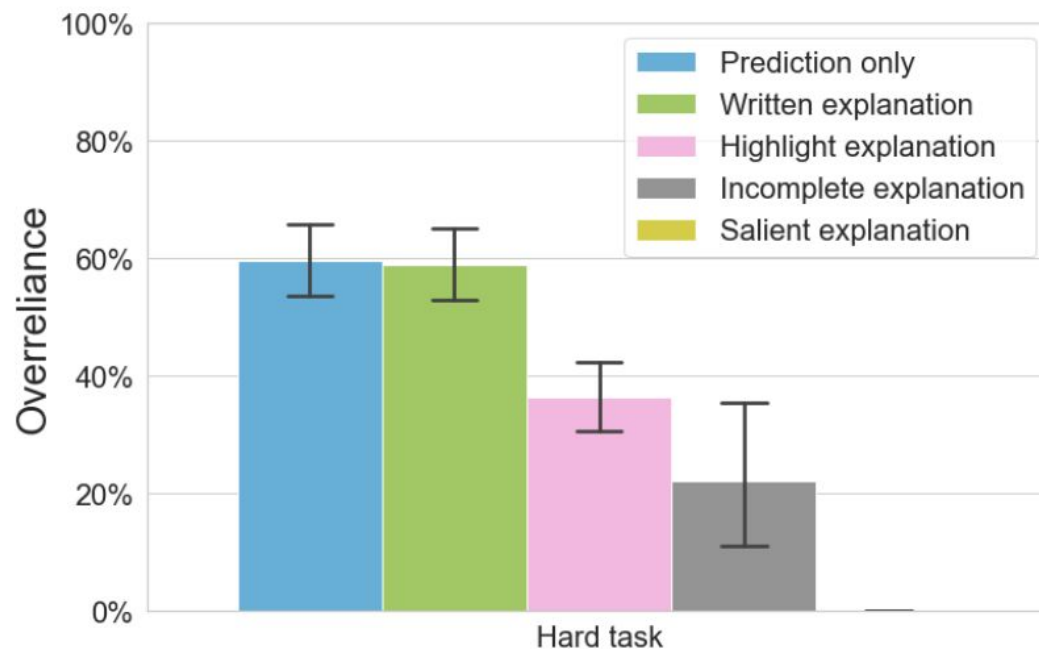
Adding two a new type of explanations:

Incomplete
explanations

Salient
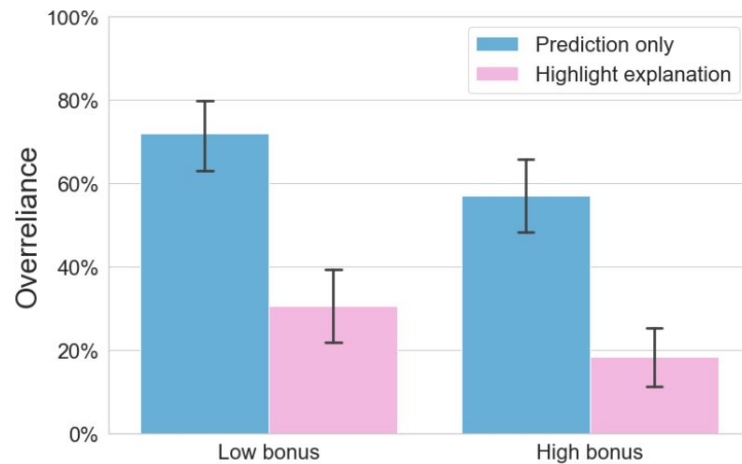explanations

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

Legend:
- Prediction only
- Written explanation
- Highlight explanation
- Incomplete explanation
- Salient explanation

Y-axis: Overreliance (0% to 100%)
X-axis: Hard task

More benefit -> less overreliance

Challenges with evaluation protocols for human-AI systems