LOST IN EXPLANATION

reflecting on interpretability desiderata with visual commonsense reasoning

Ana Marasović

Chandra Bhagavatula

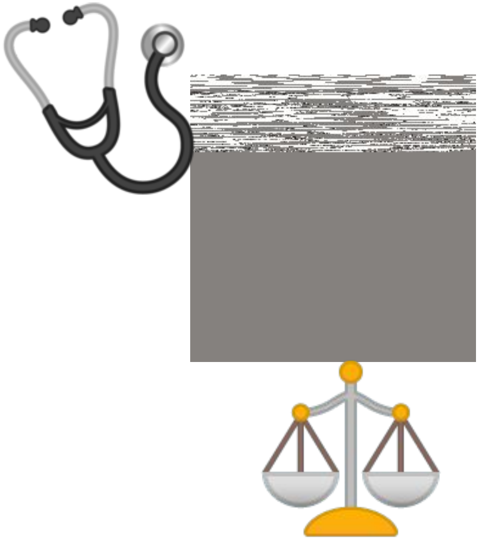
Ronan Le Bras

Yejin Choi

Allen Institute for AI

Why explainable AI?

real-world deployment



scientific methodology

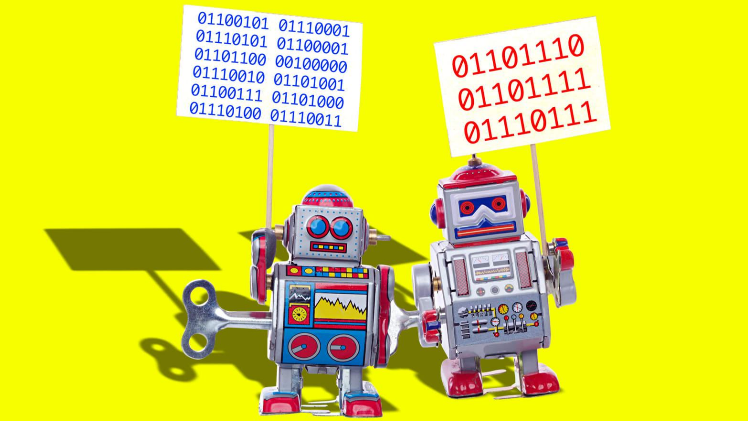
Why?

How?

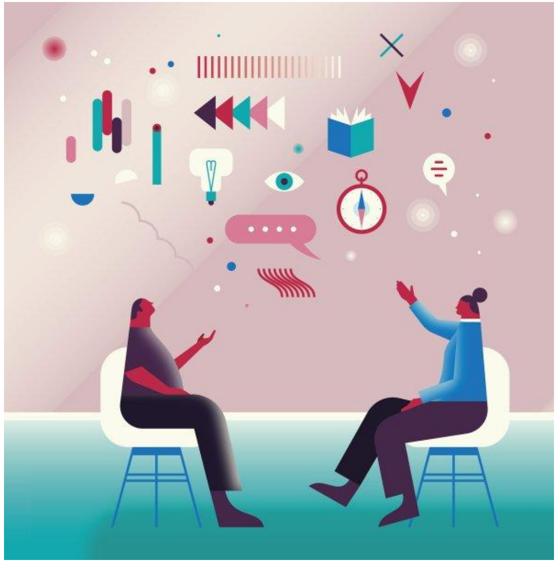
No.

What is a good explanation?

machine learning / explainable AI



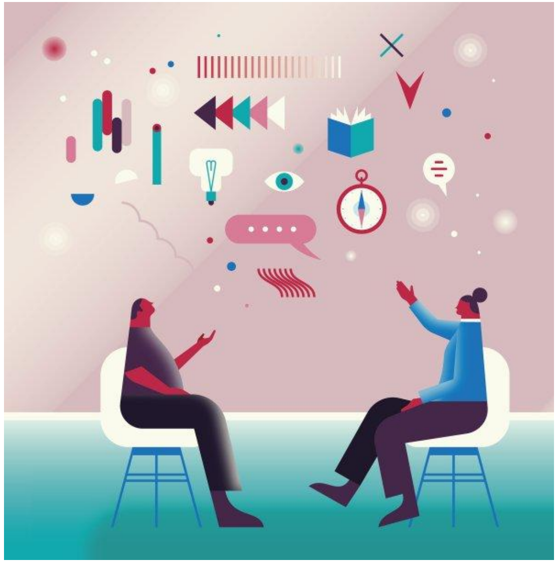
social sciences



machine learning / explainable AI

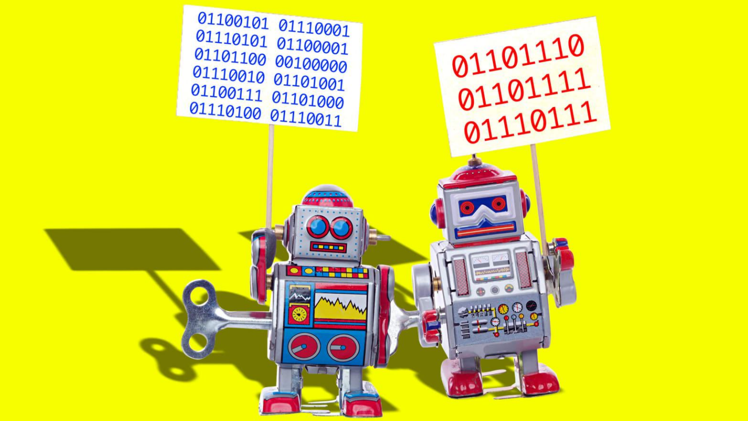
“the inmates running the asylum”

social sciences



Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. In AIJ 2018.

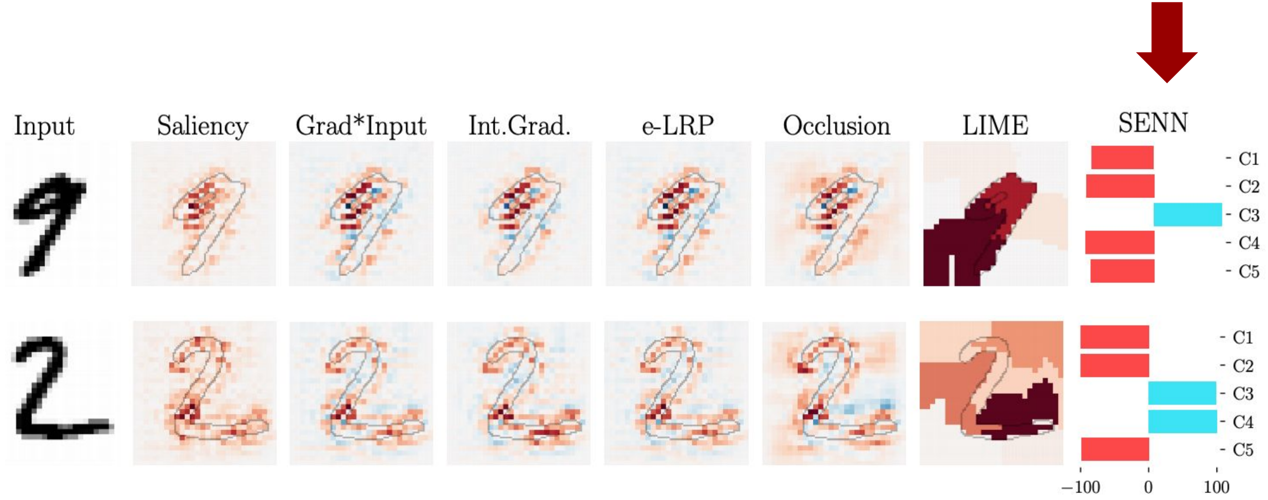
machine learning / explainable AI

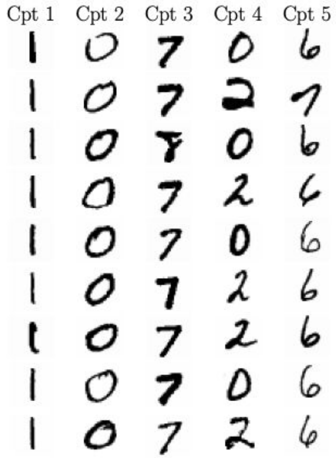


social sciences

Interpretability desiderata in machine learning

1. **Explicitness:** immediately understandable

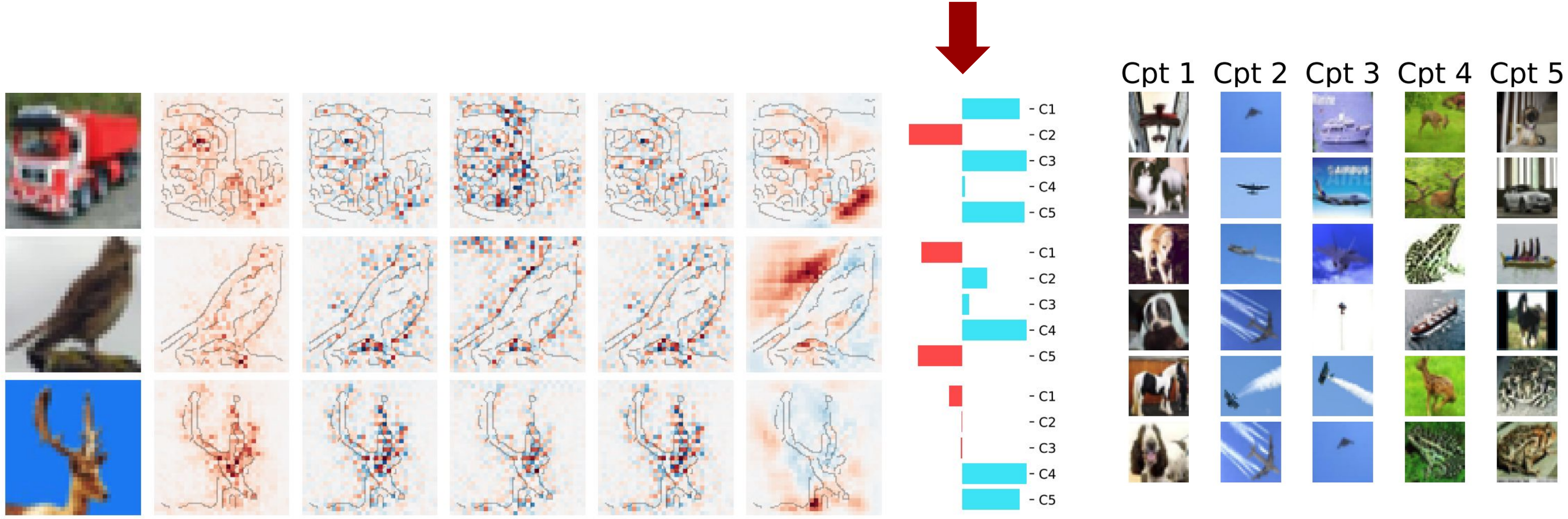




Cpt 3 = diagonal strokes
Cpt 4 = stylized 2s

Interpretability desiderata in machine learning

1. **Explicitness:** immediately understandable

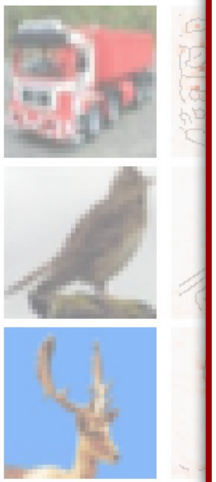


Cpt 5 = stripes and vertical lines

Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

1. **Explicitness:** immediately understandable

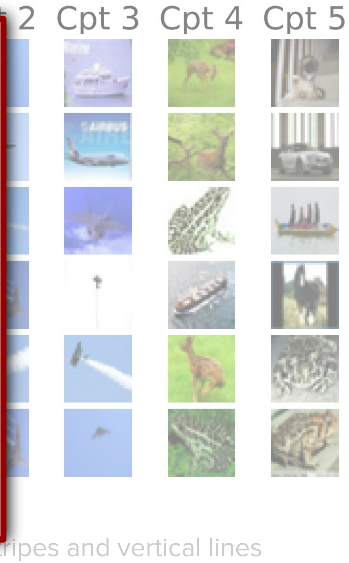


✔️ concepts instead of raw features (pixels, words)

❗ design immediately understandable concepts

⛔ author’s qualitative assessment of a few examples

❗ human evaluation

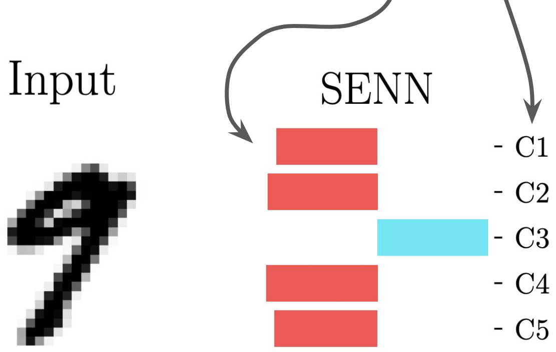


Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

2. Faithfulness: calculated relevance scores θ are “truly” relevant

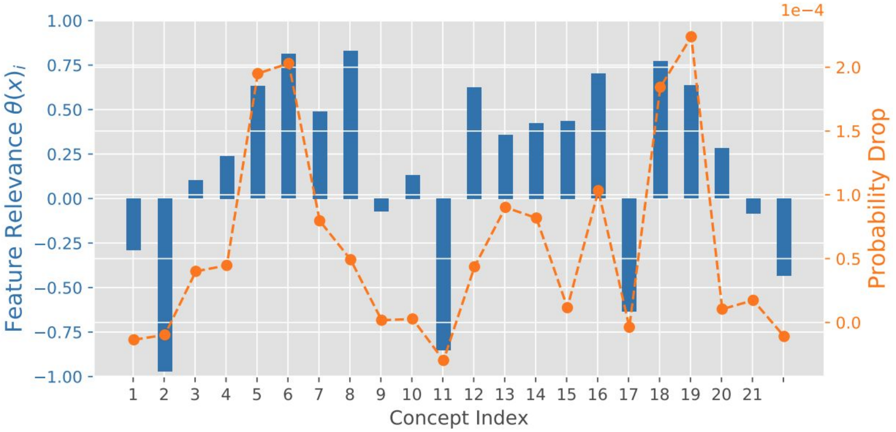
$$f(x) = \theta(x)^T h(x) = \sum_{i=1}^{K} \theta(x)_i h(x)_i$$



Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

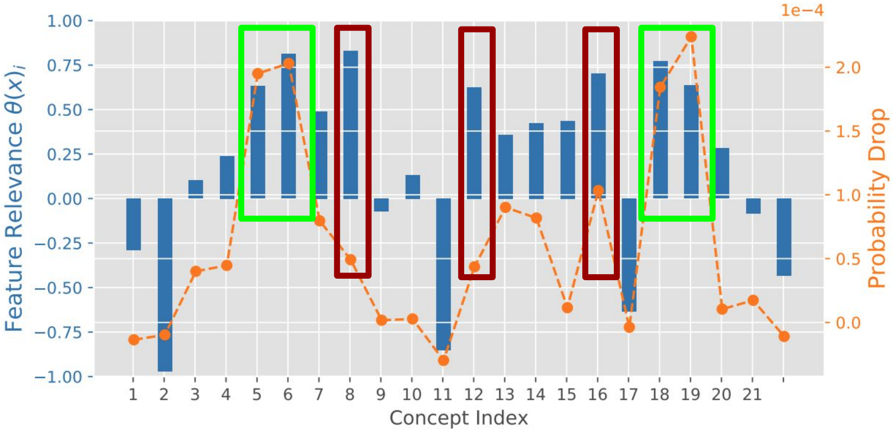
2. Faithfulness: calculated relevance scores are “true” relevance



Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

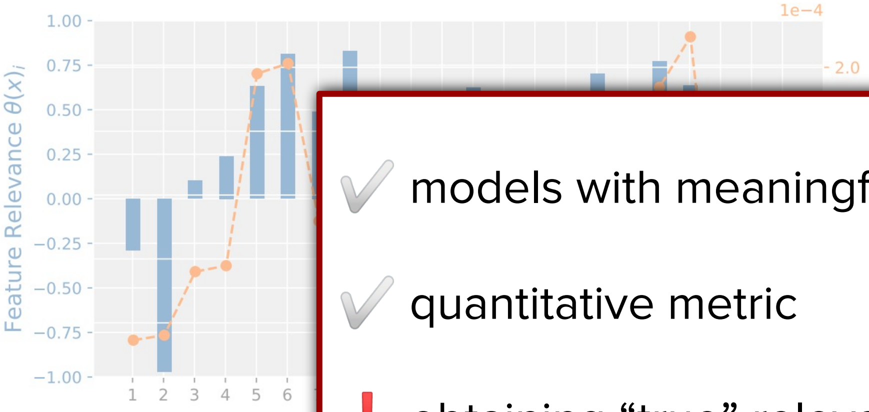
2. Faithfulness: calculated relevance scores are “true” relevance



Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

2. Faithfulness: calculated relevance scores are “true” relevance



✓ models with meaningful feature removal

✓ quantitative metric

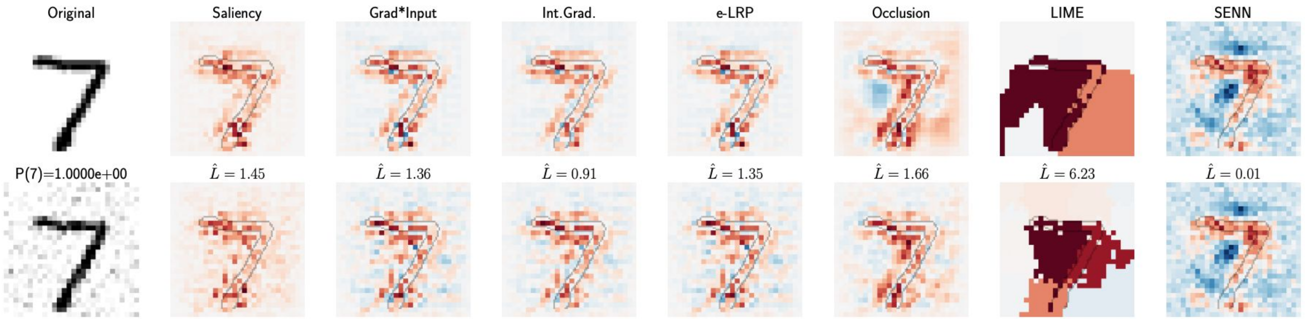
❗ obtaining “true” relevance is not trivial

Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

3. Stability: explanations are consistent for similar inputs

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f_{\text{expl}}(x_i) - f_{\text{expl}}(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

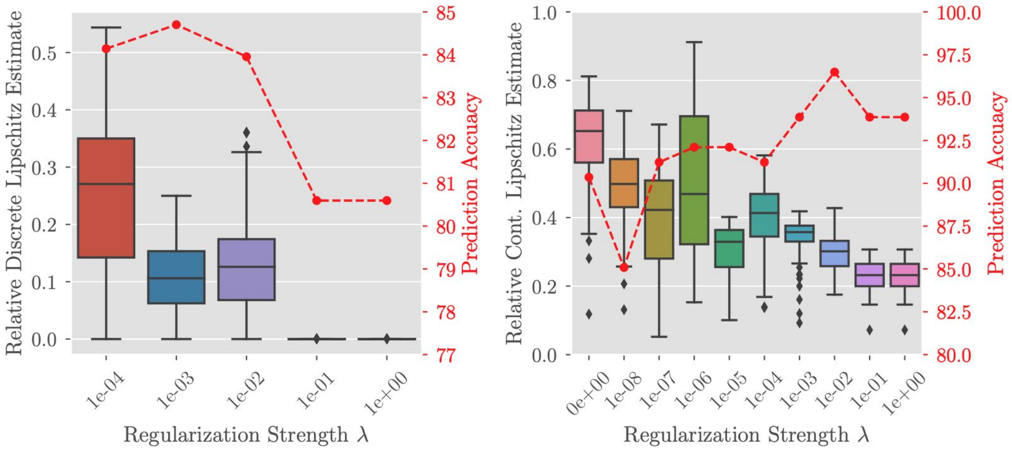


adding min. noise to the input results in visible changes in the explanations

Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

3. Stability: explanations are consistent for similar inputs



Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

Interpretability desiderata in machine learning

3. Stability: explanations are consistent for similar inputs

✔️ quantitative metric

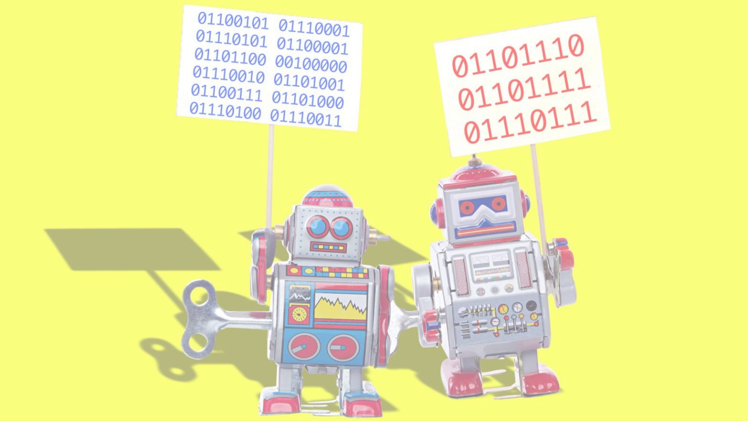
❗ interpretability approaches are not robust

❗ optimize stability of explanation

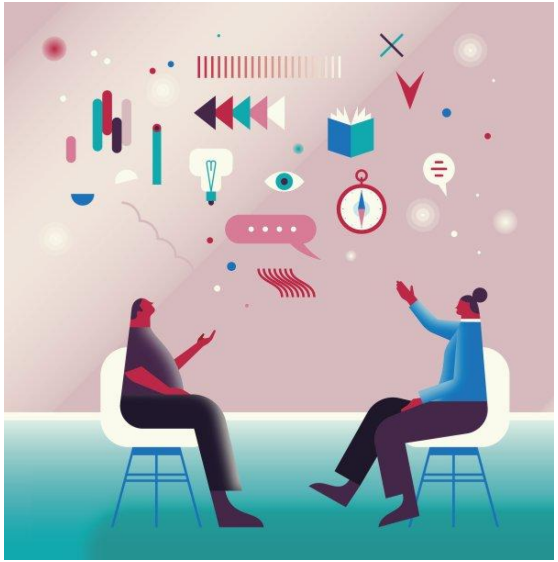
❗ tradeoff between stability and prediction accuracy

Alvarez-Melis and Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In NeurIPS 2018.

machine learning / explainable AI



social sciences



Interpretability desiderata in social science

1. Contrastive: why event happened instead of some imagined, counterfactual event?

What are the factors in the application that would need to change to get the same limit? (woman → man)

Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. In AIJ 2018.

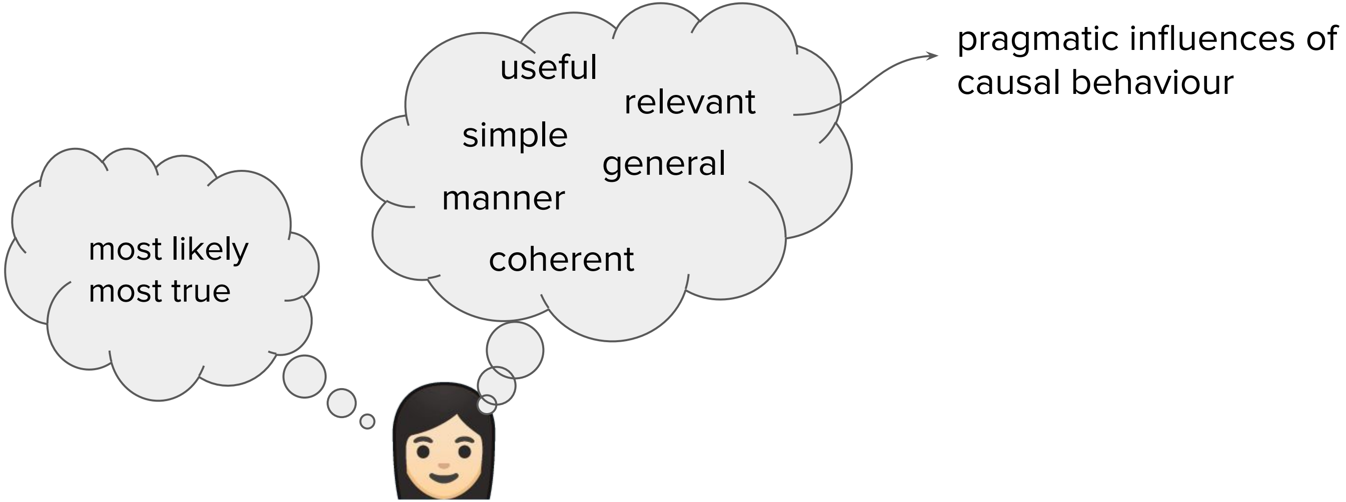
Interpretability desiderata in social science

2. Selected: explainee cares only about a small number of causes (relevant to the context)

Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. In AIJ 2018.

Interpretability desiderata in social science

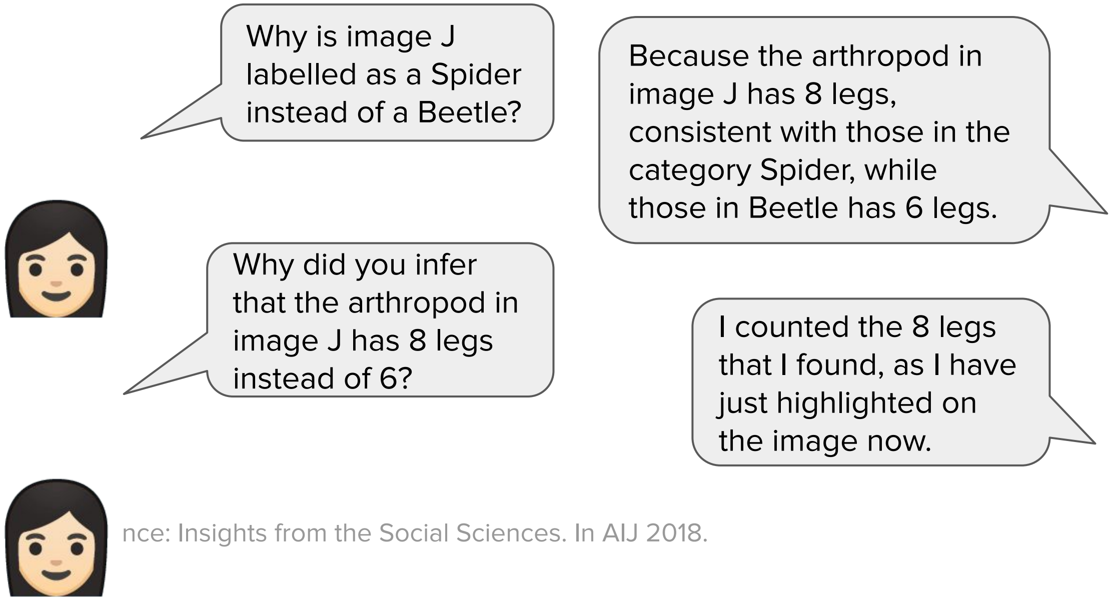
3. The most likely explanation is not always the best



Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. In AIJ 2018.

Interpretability desiderata in social science

4. Social: we interact and argue about the explanation and contextualize explanation wrt the explainee



Miller. Explanation in Artific nce: Insights from the Social Sciences. In AIJ 2018.

human evaluation

- explicitness
- usefulness
- relevance
- simplicity
- coherence
- rules of conversation

explanation evaluation

automatic evaluation

- faithfulness
- stability

model design

- concepts instead of raw inputs
- understandable concepts
- meaningful feature removal
- small number of causes
- contrast with counterfactual
- interactive conversations

optimization

- stability

explanation generation and selection

human evaluation

explicitness
usefulne
relevan
simplic
coherer
rules of conv

automatic evaluation

faithfulness

ign

of raw inputs
concepts
re removal
small number of causes
contrast with counterfactual
interactive conversations

optimi

stability

design, optimize, and evaluate compositional self-explanatory reasoning models

Visual Commonsense Reasoning (VCR)

ideas and challenges

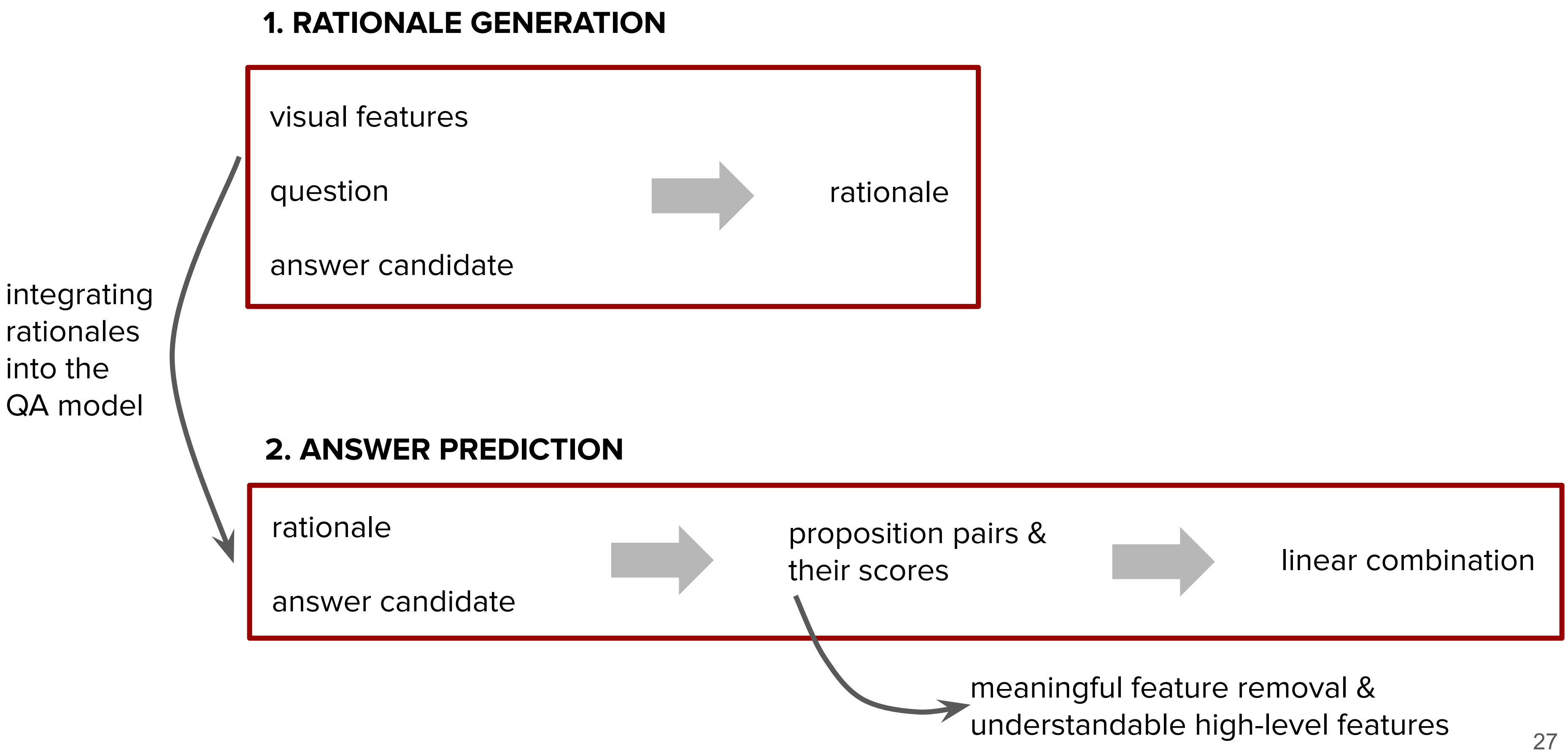
“Given a challenging question about an image, a machine must answer correctly and then provide a rationale justifying its answer.”

https://visualcommonsense.com/

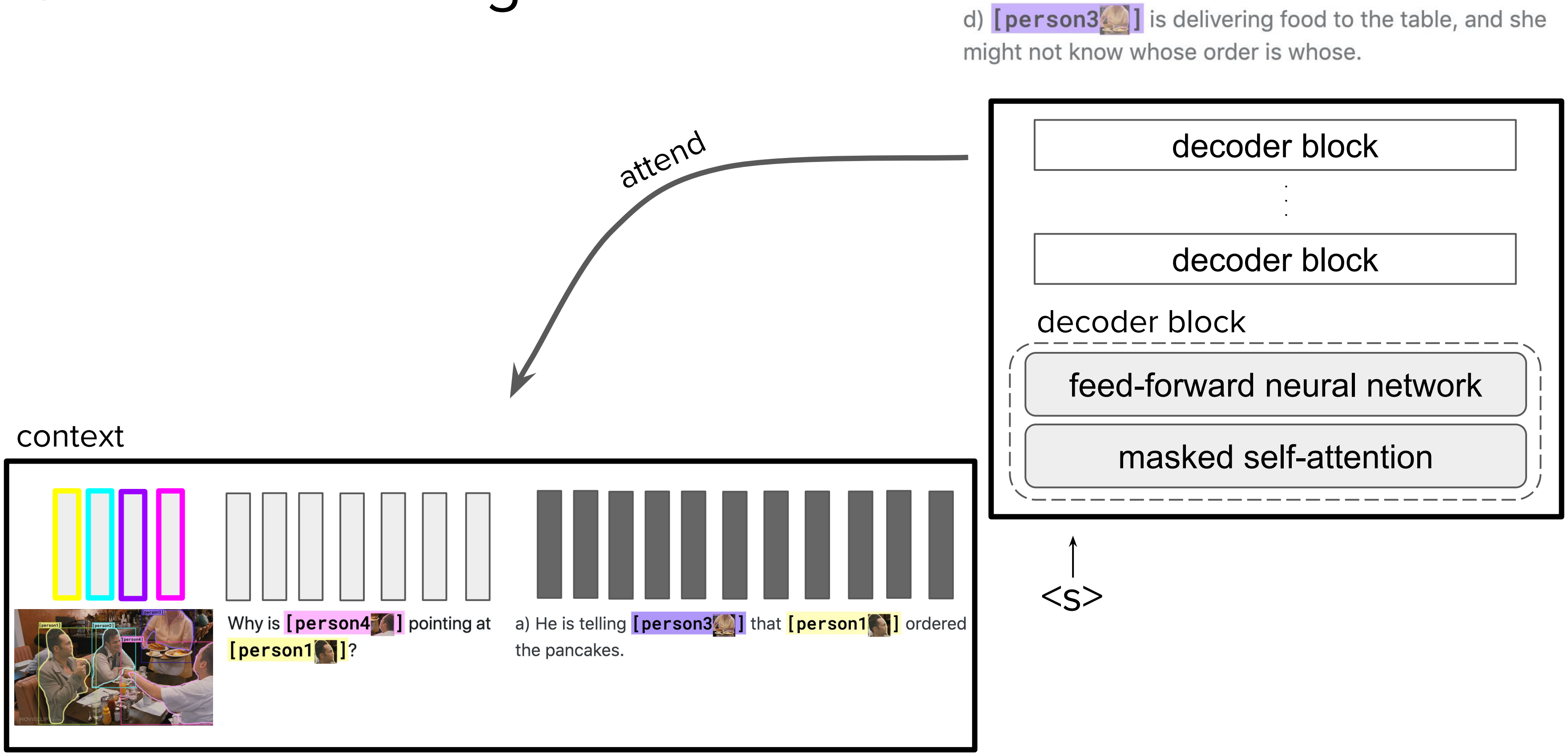
👍 VCR requires **cognition-level reasoning** (inferring the likely intents, goals, and social dynamics of people)

🤔 **Are models that correctly classify 4 rationale choices really justifying their answer prediction?**

💡 Design a model where the **rationale is intrinsic to the model**...
... and do not forget explainability desiderata


1. RATIONALE GENERATION
visual features
question
answer candidate
rationale
integrating rationales into the QA model
2. ANSWER PREDICTION
rationale
answer candidate
proposition pairs & their scores
linear combination
meaningful feature removal & understandable high-level features

GPT-2 rationale generation



Proxy for generation evaluation

[CLS] **gold** rationale

[SEP] answer candidate 1 [SEP]
[SEP] answer candidate 2 [SEP]
[SEP] answer candidate 3 [SEP]
[SEP] answer candidate 4 [SEP]

*actually RoBERTa

89.28%

DROP?

[CLS] **generated** rationale

[SEP] answer candidate 1 [SEP]
[SEP] answer candidate 2 [SEP]
[SEP] answer candidate 3 [SEP]
[SEP] answer candidate 4 [SEP]

?

Non-compositional answer prediction

[SEP] generated rationale 1 [SEP] answer candidate 1 [SEP]
[SEP] generated rationale 2 [SEP] answer candidate 2 [SEP]
[SEP] generated rationale 3 [SEP] answer candidate 3 [SEP]
[SEP] generated rationale 4 [SEP] answer candidate 4 [SEP]

*actually RoBERTa

$a \in \{1, 2, 3, 4\}$

What are concepts?

d) [person3] is delivering food to the table, and she might not know whose order is whose.

a) He is telling [person3] that [person1] ordered the pancakes.

Too many words + not "high-level features"
How about propositions?

Compositional (?) answer prediction

“generated” rationale

d) [person3] is delivering food to the table, and she might not know whose order is whose.

candidate answer

a) He is telling [person3] that [person1] ordered the pancakes.

Compositional (?) answer prediction

“generated” rationale

d) [person3] is delivering food to the table, and she might not know whose order is whose.

PredPatt

observation representations

P3 is delivering food to the table

she might not know whose order is whose

candidate answer

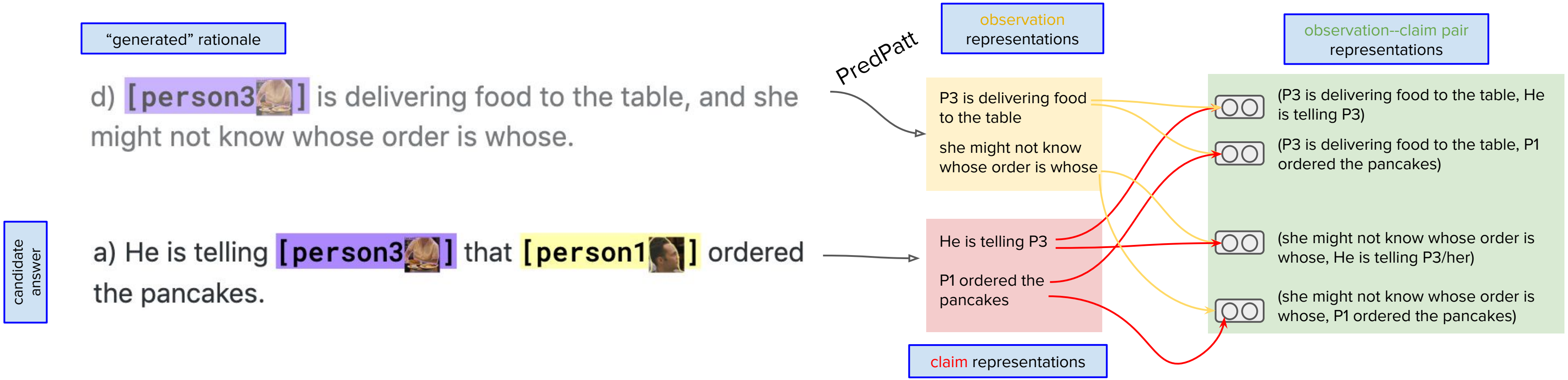
a) He is telling [person3] that [person1] ordered the pancakes.

He is telling P3

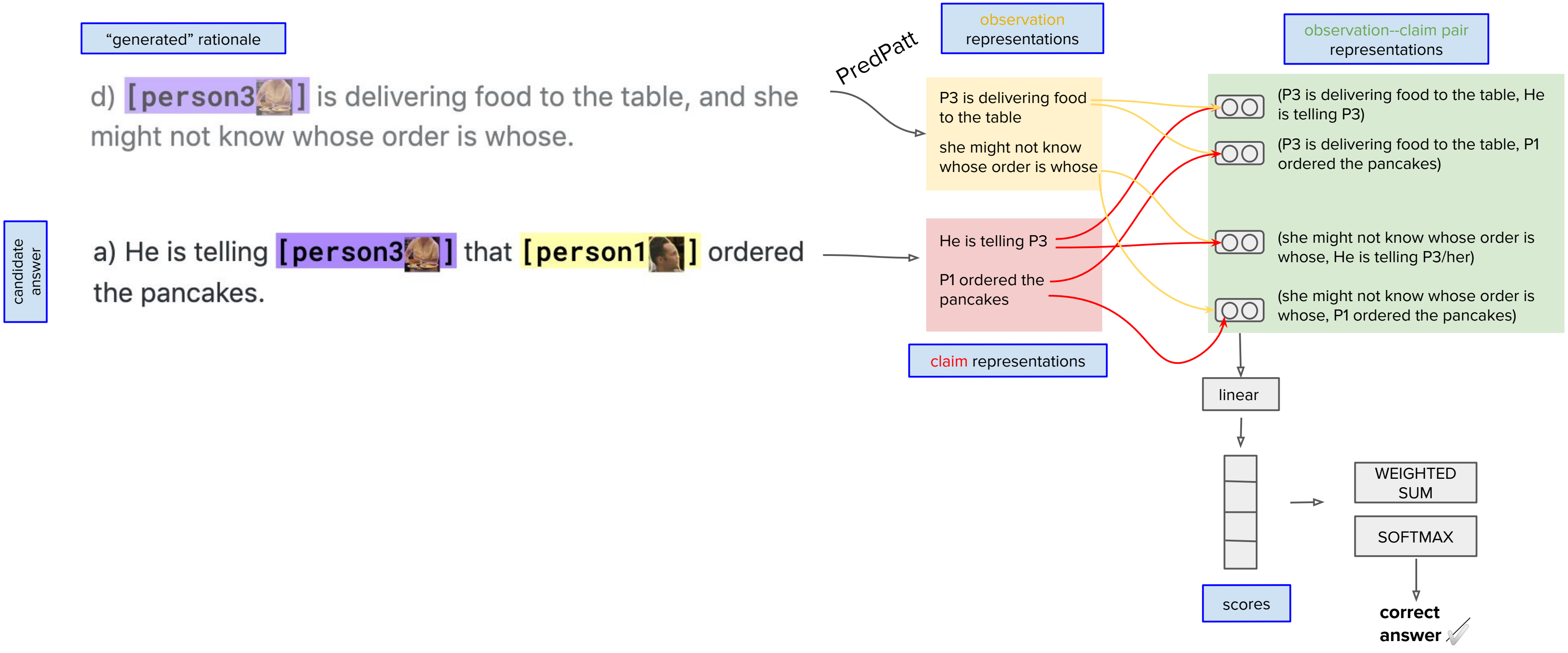
P1 ordered the pancakes

claim representations

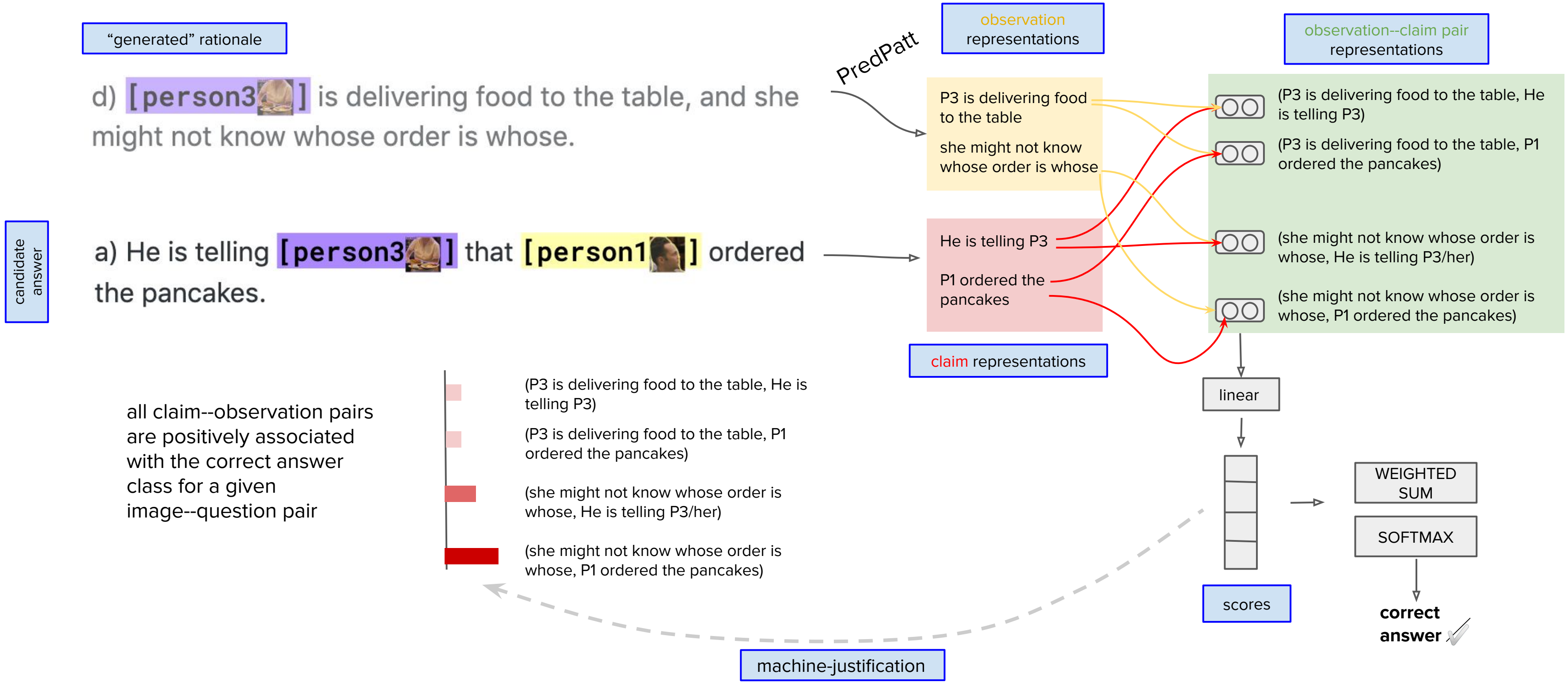
Compositional (?) answer prediction



Compositional (?) answer prediction



Compositional (?) answer prediction



Challenge #1: predicate-argument extraction

rationale
they are here together , look similar , and have an age disparity .

current propositions (by PredPatt) ✅
they are here together
they look similar
they have an age disparity

Sheng et al. An Evaluation of PredPatt and Open IE via Stage 1 Semantic Role Labeling. In IWCS 2017.
White et al. Universal Decompositional Semantics on Universal Dependencies. In EMNLP 2016.

Challenge #1: predicate-argument extraction

rationale
cabs usually wait for people to get in *before* they pull away

current propositions (by PredPatt) ✗
cabs usually wait for people to get in
they pull away

wanted proposition ✅
cabs (usually) wait for people to get in before they pull away

Challenge #1: predicate-argument extraction

rationale
jessie is dressed in less fancy clothing indicating that they are a squire . riley is climbing up to the top of horse jessie is in position to steady the horse .

current propositions (by PredPatt) ✗
jessie is dressed in less fancy clothing
indicating they are a squire
they are a squire

wanted proposition ?
jessie is dressed in less fancy clothing
they are a squire

Challenge #2: What if a wrong answer is justified well?

1. Why is [person5] smiling?

a) Because she is happy about [person5] blowing a horn.	0.0%
b) [person5] is anticipating her soon to occur wedding and is happy about it.	2.0%
c) [person5] is smiling because she is helping someone.	59.4%
d) [person5] is showing love to her friend.	38.6%

a generated rationale might make sense when you read it…
… but a horn still won't be visible on the photo

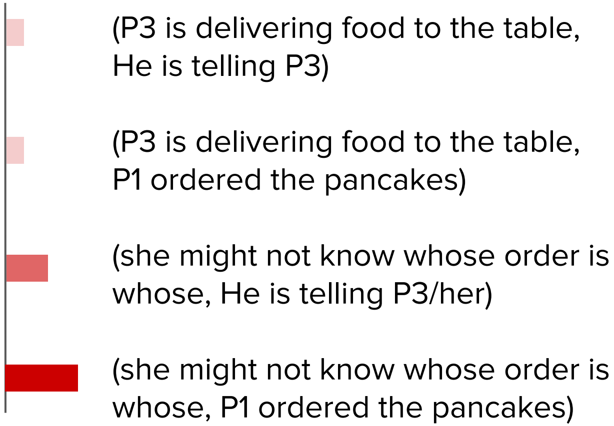
A special ingredient: discriminator

a) Because she is happy about [person5] blowing a horn.

Tan and Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In EMNLP 2019.

Kim et al. Image Captioning with Very Scarce Supervised Data: Adversarial Semi-Supervised Learning Approach. In EMNLP 2019.

Final machine-justification



+

image-rationale pair do not contradict

image-answer candidate pair do not contradict

human evaluation

explicitness
usefulness
relevance
simplicity
coherence
rules of conversation

automatic evaluation

faithfulness

stability

explanation evaluation

model design

concepts instead of raw inputs

understandable concepts

meaningful feature removal

small number of causes

contrast with counterfactual

interactive conversations

optimization

stability

explanation generation and selection

Some future ideas…

human evaluation

explicitness
usefulness
relevance
simplicity
coherence
rules of conversation

automatic evaluation

faithfulness

stability

explanation evaluation

model design

concepts instead of raw inputs

understandable concepts

meaningful feature removal

small number of causes

contrast with counterfactual

interactive conversations

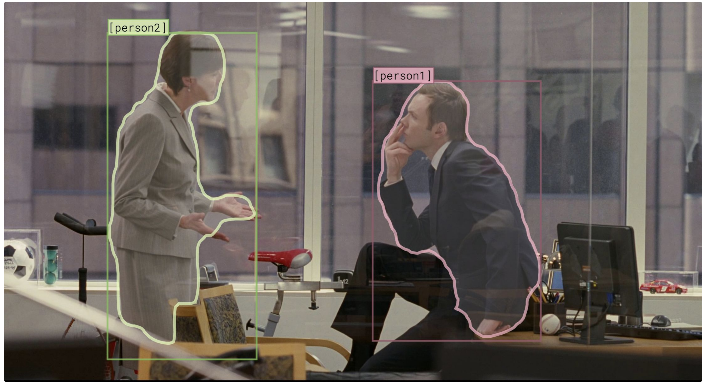
optimization

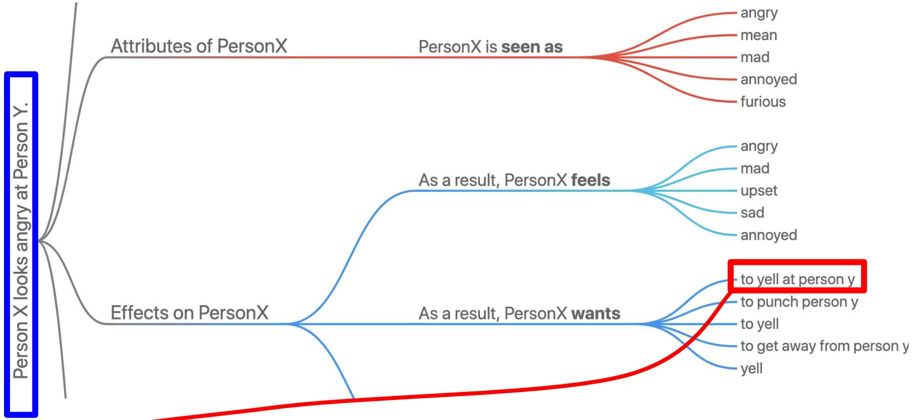
stability

explanation generation and selection

Inducing "social" biases

https://mosaickg.apps.allenai.org/





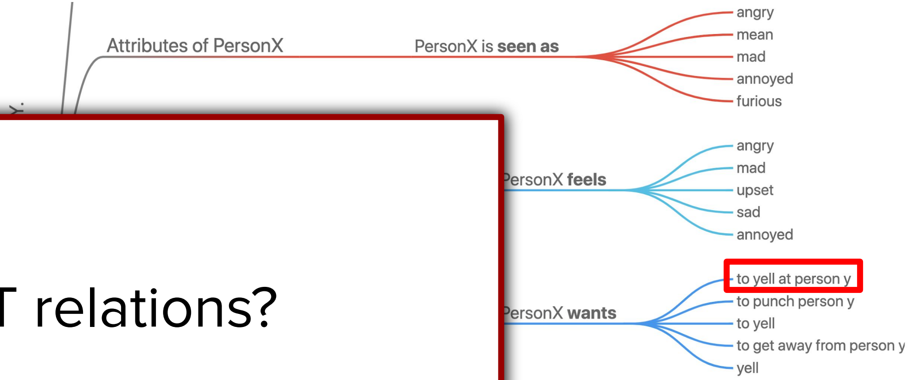
1. What is **[person2]** doing?

a) She is going to spike the punch in **[sportsball1]** .	**0.0%**
b) She is yelling at **[person1]** .	**76.2%**
c) She is checking out of the place.	**23.3%**
d) She is licking her lips.	**0.4%**

I think so because...

a) Her hands are held up in frustration and she looks angry at **[person1]** .	**64.8%**
b) She has an angry look and has her hands cupped around her mouth to be louder.	**30.4%**
c) She is storming away from him, and his he pleading with her.	**2.0%**
d) She speaks softly so that only **[person1]** can hear what she is saying.	**2.7%**

Inducing “social” biases



Which COMET relations?

For which examples?

1. What is [person

a) She is going to sp
b) She is yelling at [
c) She is checking out of the place. **23.3%**
d) She is licking her lips. **0.4%**

she looks angry at [person1]
s cupped around her mouth to be louder. **30.4%**
c) She is storming away from him, and his he pleading with her. **2.0%**
d) She speaks softly so that only [person1] can hear what she is saying. **2.7%**

human evaluation

explicitness
usefulness
relevance
simplicity
coherence
rules of conversation

automatic evaluation

faithfulness

stability

explanation evaluation

model design

concepts instead of raw inputs

understandable concepts

meaningful feature removal

small number of causes

contrast with counterfactual

interactive conversations

optimization

stability

explanation generation and selection

Multi-modal explanations

IMO pointing to his face is more understandable than describing it

2. Does [person2] enjoy [person1] 's singing?

a) No, [person2] is not happy. **0.0%**
b) No, [person2] does not know the words to the song. **0.0%**
c) Yes, [person2] is tired of [person1] 's rebellious attitude. **0.0%**
d) Yes, [person2] enjoy's [person1] 's singing. **100.0%**

I think so because...

a) [person2] is sitting in [couch1] and has his eyes on [person1] . **1.0%**
b) [person2] is giving [person1] his full attention, with his head tilted to better listen and his eyes focused exclusively on [person1] . **3.8%**
c) [person2] plays his instrument with passion as the look on his face is of pure excitement. **0.2%**
d) [person2] looks very relaxed with his eyes closed and his face resting on his hand. **95.0%**

Park et al. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In CVPR 2018.

Multi-modal explanations

She does not know whose order is whose.

IMO textual rationale is more understandable

Park et al. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In CVPR 2018.

now evaluate my
(human) explanations :)

thank you!

https://github.com/amarasovic/interpretability-literature/

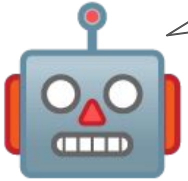
(generic / squashed models)

human rationale

machine justification

Premise: Three dogs racing on racetrack.

Hypothesis: Three **cats** race on a track.



Gururangan et al. Annotation Artifacts in Natural Language Inference Data. In NAACL 2018.

(compositional / modular / transparent models)

human rationale

machine justification

machine learning
(Alvarez-Melis and Jaakkola, 2018)

✅ **Explicit:**
immediately understandable

✅ **Faithful:**
calculated relevance scores are “true” relevance

✅ **Stable:**
explanations are consistent for similar inputs

social science
(Miller, 2018)

❌ **Contrastive:**
why event happened instead of some imagined, counterfactual event

✅ **Selected:**
explainee cares only about a small number of causes of an event (relevant to the context)

✅ **The most likely explanation is not always the best**

❌ **Social:**
we interact and argue about the explanation and contextualize explanation wrt the explainee