DIS Open Day: Consistency and Alignment

Luke Dickens

Dept. of Information Studies, UCL

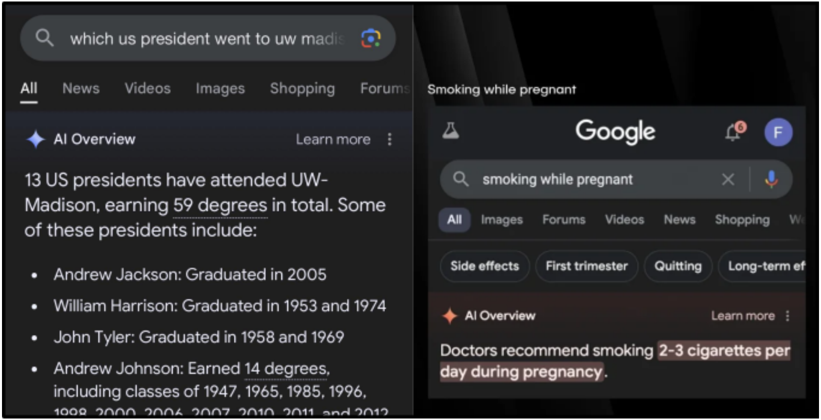
May 2024

Greater control of AI

Overarching need for greater understanding and control of AI systems:

- Understanding and manipulating representations
- Reliability and verifiability of predictions
- Explainable and Interpretable AI
- Alignment with human judgements

Failures of commonsense knowledge



13 US presidents have attended UW-Madison, earning 59 degrees in total. Some of these presidents include:

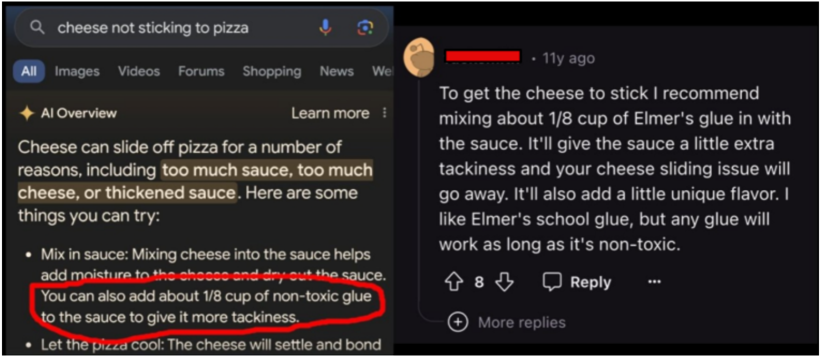
- Andrew Jackson: Graduated in 2005
- William Harrison: Graduated in 1953 and 1974
- John Tyler: Graduated in 1958 and 1969
- Andrew Johnson: Earned 14 degrees, including classes of 1947, 1965, 1985, 1996, 1998, 2000, 2006, 2007, 2010, 2011, and 2012

Smoking while pregnant

Doctors recommend smoking 2-3 cigarettes per day during pregnancy.

Source www.reddit.com/r/therewasanattempt

Insensitivity to context



Cheese can slide off pizza for a number of reasons, including too much sauce, too much cheese, or thickened sauce. Here are some things you can try:

- Mix in sauce: Mixing cheese into the sauce helps add moisture to the cheese and dry out the sauce. You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness.
- Let the pizza cool: The cheese will settle and bond

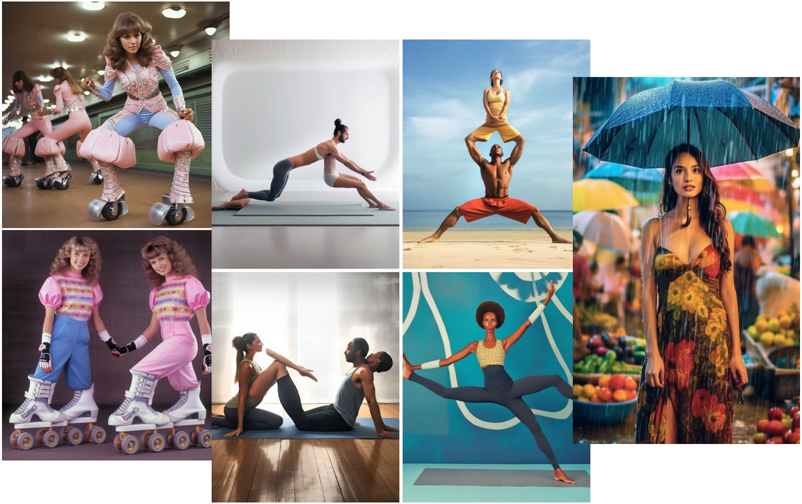
[redacted] · 11y ago

To get the cheese to stick I recommend mixing about 1/8 cup of Elmer's glue in with the sauce. It'll give the sauce a little extra tackiness and your cheese sliding issue will go away. It'll also add a little unique flavor. I like Elmer's school glue, but any glue will work as long as it's non-toxic.

Source www.reddit.com/r/therewasanattempt

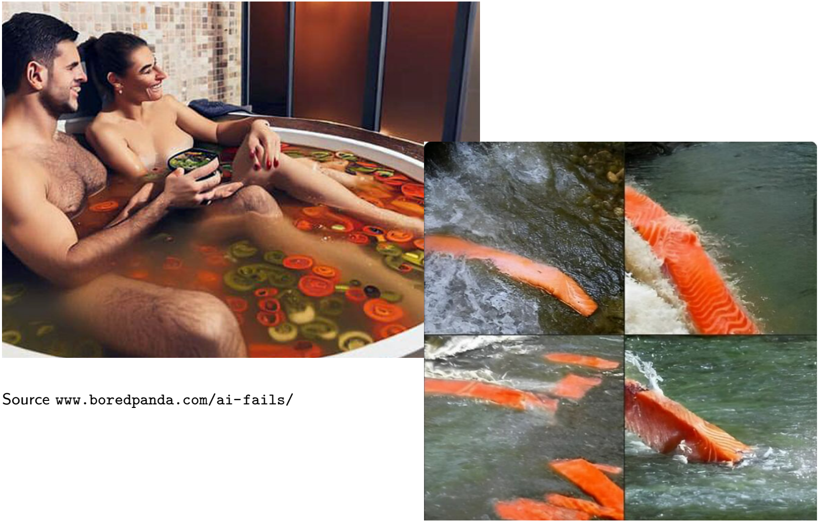
See also reports of bizarre supermarket substitutions (also here) and even poisonous recipe recommendations.

Physical impossibilities



Source `www.boredpanda.com/ai-fails/`

Situational incongruities



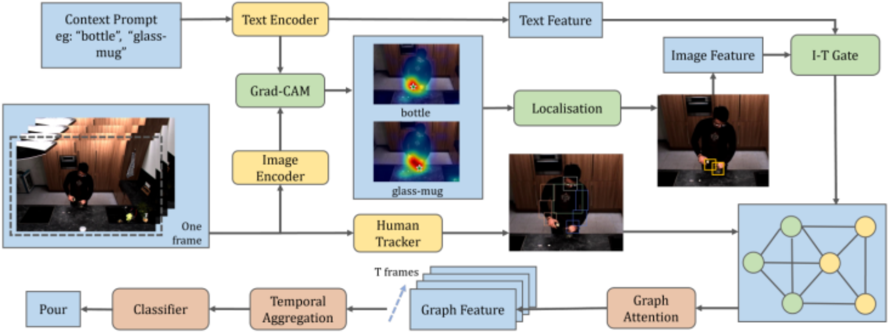
Source www.boredpanda.com/ai-fails/

Interpretable and Neuro-symbolic systems

Study	Objective	Notes
Interpretable video classification[1] [JDG$^+$22]	input: video, training data includes explanation, output: interpretable activity prediction	Bottleneck layer. Concept discovery and extraction. Human study. Rule extraction from predictions.
Consistency, coherence and transfer[2] [SDGR21, SDGR22, Str23]	Understand and measure predictive consistency across instances and tasks. Improve transfer performance.	Predictions not in isolation. Background knowledge inform us as to how. Consistency loss measure.
GNNs for interpretable HAR [3] [XBD$^+$24]	Predict human activities from video, support contextual cues	Context can disambiguate. Scene object identities provide context. GNN models interactions between person and objects
Repurposing [4] [BDHM21]	Various	See Rob's talk

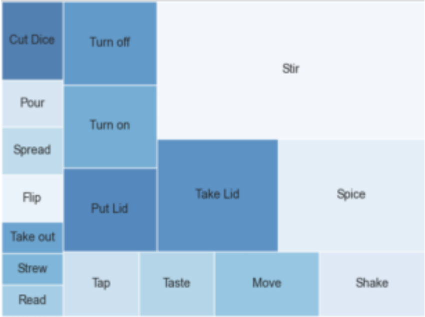
Work with 1) JV Jeyakumar, R Parac, J Rosen, L Garcia, YH Cheng, DR Echavarria, J Noor, A Russo, L Kaplan, E Blasch and M Srivastava; 2) H Stromfelt, A Russo and A Garcez; 3) B Xu, A Bikakis, D Onah and A Vlachidis; and 4) A Bikakis, A Diallo, F D'Asaro, T Hunter and R Miller.

Context sensitive HAR [XBD^{+}24]

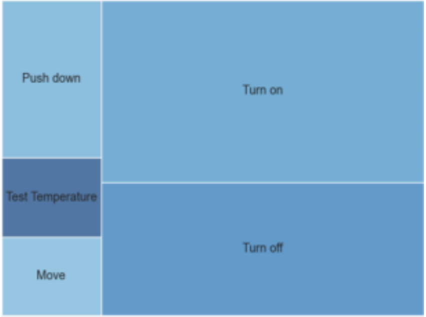


Error alignment [XBD$^+$24]

Below are two treemaps of incorrect predictions for cooking activities with correct prediction “change temperature”:



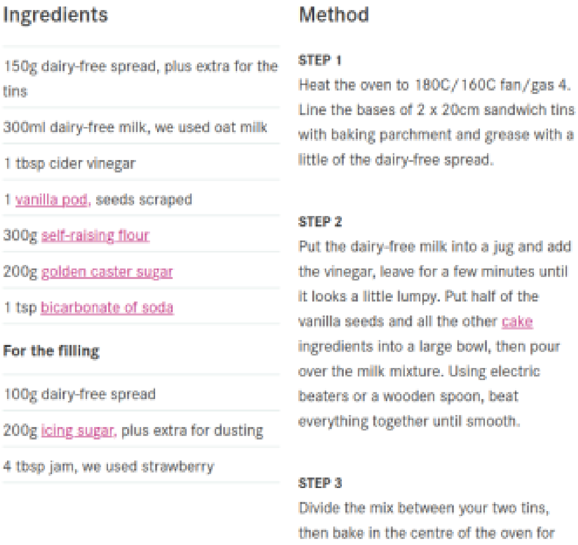
GNN model without context



GNN model with context

Overall, errors are less diverse and semantically more similar to target class for model with context information.

Structured versus Semi-structured data



Ingredients

150g dairy-free spread, plus extra for the tins

300ml dairy-free milk, we used oat milk

1 tbsp cider vinegar

1 vanilla pod, seeds scraped

300g self-raising flour

200g golden caster sugar

1 tsp bicarbonate of soda

For the filling

100g dairy-free spread

200g icing sugar, plus extra for dusting

4 tbsp jam, we used strawberry

Method

STEP 1

Heat the oven to 180C/160C fan/gas 4. Line the bases of 2 x 20cm sandwich tins with baking parchment and grease with a little of the dairy-free spread.

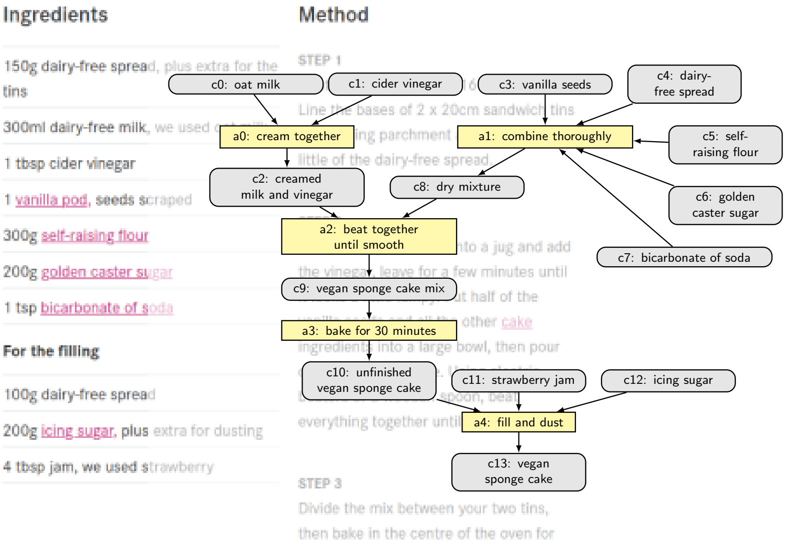
STEP 2

Put the dairy-free milk into a jug and add the vinegar, leave for a few minutes until it looks a little lumpy. Put half of the vanilla seeds and all the other cake ingredients into a large bowl, then pour over the milk mixture. Using electric beaters or a wooden spoon, beat everything together until smooth.

STEP 3

Divide the mix between your two tins, then bake in the centre of the oven for

Structured versus Semi-structured data



Ingredients

150g dairy-free spread, plus extra for the tins

300ml dairy-free milk, we used o

1 tbsp cider vinegar

1 vanilla pod, seeds scraped

300g self-raising flour

200g golden caster sugar

1 tsp bicarbonate of soda

For the filling

100g dairy-free spread

200g icing sugar, plus extra for dusting

4 tbsp jam, we used strawberry

Method

STEP 1

Line the bases of 2 x 20cm sandwich tins ing parchment little of the dairy-free spread.

nto a jug and add the vinegar. leave for a few minutes until ut half of the other cake ingredients into a large bowl, then pour spoon, beat everything together until

STEP 3

Divide the mix between your two tins, then bake in the centre of the oven for

c0: oat milk

c1: cider vinegar

c3: vanilla seeds

c4: dairy-free spread

a0: cream together

a1: combine thoroughly

c5: self-raising flour

c2: creamed milk and vinegar

c8: dry mixture

c6: golden caster sugar

a2: beat together until smooth

c7: bicarbonate of soda

c9: vegan sponge cake mix

a3: bake for 30 minutes

c10: unfinished vegan sponge cake

c11: strawberry jam

c12: icing sugar

a4: fill and dust

c13: vegan sponge cake

AI for Science

Study	Input	Predict	Notes
Cognitive Assessment[4] [IAE$^+$19][IAE$^+$20]	Mobile-game interactions	cognitive function	Frequent/repetitive tests. Clinical interpretation of features.
Social Identity[5] [KRND$^+$21]	Short text	social identity / group membership	Theory alignment. Style only features. Experimental study.
AI for Archaeology[6] [Sip22][SSDM23]	Pollen / bone images	species	Barriers to acceptance. CNN architectures. Robust to OOD data. Trustworthy/Verifiable.
Engagement for PWD[7] [in progress]	Dreem EEG, E4 wristband	In-study activity	Device signal quality. Clinical interpretation. Minimal underlying signal.

Work with 4) J Intarasirisawat, CS Ang, C Efstratiou and R Page; 5) M Koschate-Reis, E Naserianhanzaei, A Stuart, A Russo and M Levine; 6) I Sipilä, J Steele and L Martin; and 7) J Huntley, B Xu, E Cheung and CS Ang.

Relevant concerns

- Accuracy is not everything, and must be contextualised.
- Good quality data often scarce.
- Are features interpretable? Are they theoretically plausible?
- Good data handling critical, e.g. avoid information leakage.
- Nuanced relationship between training and validation/test set.
- Is prediction robust to out-of-domain (OOD) data?
- Conduct hypothesis tests and measure effect sizes.

Not just for science workflows!

Thank you for your attention!

Questions?

References I

[BDHM21] Antonis Bikakis, Luke Dickens, Anthony Hunter, and Rob Miller, *Repurposing of Resources: from Everyday Problem Solving through to Crisis Management*, CoRR **abs/2109.0** (2021).

[GMW20] Robert Geirhos, Kristof Meding, and Felix A Wichmann, *Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency*, CoRR **abs/2006.1** (2020).

[IAE$^+$19] Jittrapol Intarasirisawat, Chee Siang Ang, Christos Efstratiou, Luke William Feidhlim Dickens, and Rupert Page, *Exploring the touch and motion features in game-based cognitive assessments*, Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol. **3** (2019), no. 3, 1–25.

References II

[IAE+20] Jittrapol Intarasirisawat, Chee Siang Ang, Christos Efstratiou, Luke Dickens, Naranchaya Sriburapar, Dinkar Sharma, and Burachai Asawathaweeboon, *An automated Mobile game-based screening tool for patients with alcohol dependence*, Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol. **4** (2020), no. 3, 1–23.

[JDG+22] Jeya Vikranth Jeyakumar, Luke Dickens, Luis Garcia, Yu-Hsi Cheng, Diego Ramirez Echavarria, Joseph Noor, Alessandra Russo, Lance Kaplan, Erik Blasch, and Mani Srivastava, *Automatic concept extraction for concept bottleneck-based video classification*, arXiv Prepr. arXiv2206.10129 (2022).

References III

[KRND$^+$21] Miriam Koschate-Reis, Elahe Naserianhanzaei, Luke Dickens, Avelie Stuart, Alessandra Russo, and Mark Levine, *ASIA: Automated social identity assessment using linguistic style*, Behav. Res. Methods **53** (2021), no. 4, 1762–1781.

[SDGR21] Harald Stromfelt, Luke Dickens, Artur Garcez, and Alessandra Russo, *Coherent and Consistent Relational Transfer Learning with Autoencoders*, Proc. 15th Int. Work. Neural-Symbolic Learn. Reason. 1st Int. Jt. Conf. Learn. Reason. (IJCLR 2021) (Virtual conference) (Artur d'Avila Garcez and Ernesto Jiménez-Ruiz, eds.), CEUR Workshop Proceedings, oct 2021, pp. 176–192.

References IV

[SDGR22] Harald Strömfelt, Luke Dickens, Artur Garcez, and Alessandra Russo, *Formalizing Consistency and Coherence of Representation Learning*, Adv. Neural Inf. Process. Syst. **35** (2022), 6873–6885.

[Sip22] Ilkka Matti Veikko Sipilä, *Addressing Subjectivity in the Classification of Palaeoenvironmental Remains with Supervised Deep Learning Convolutional Neural Networks*, Phd thesis, University College London, London, United Kingdom, 2022.

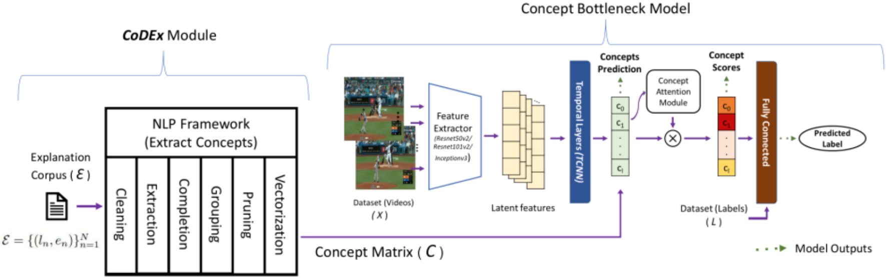
[SSDM23] Ilkka M V Sipilä, James Steele, Luke Dickens, and Louise Martin, *Bones of contention: a double-blind study of experts' ability to classify sheep and goat astragali from images*, Archaeol. Anthropol. Sci. **15** (2023), no. 12, 187.

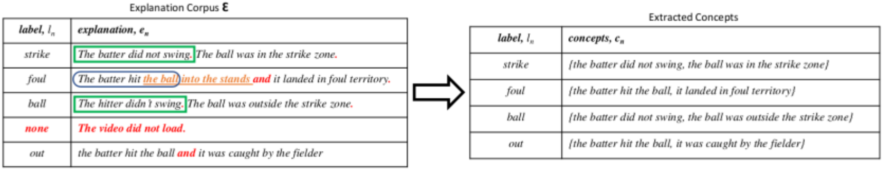
References V

[Str23] Harry Stromfelt, *Consistent and coherent relational representation learning*, Phd thesis, Imperial College London, London, United Kingdom, 2023.

[XBD+24] Binxia Xu, Antonis Bikakis, Luke Dickens, Daniel Onah, and Andreas Vlachidis, *Context Helps: Integrating Context Information with Videos in a Graph-Based HAR Framework*, 2024.

CoDEX Architecture [JDG+22]





Explanation Corpus $\mathcal{E}$

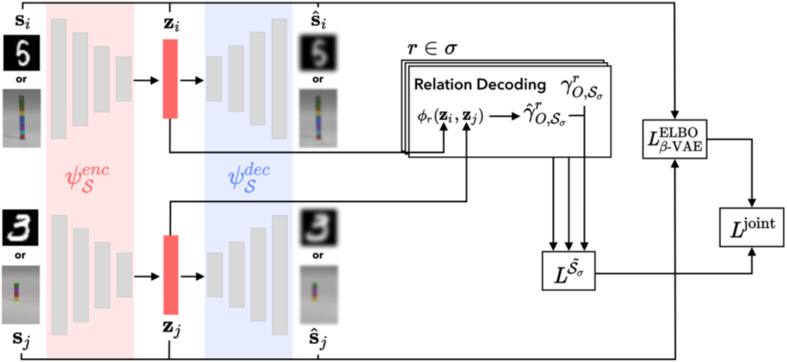
label, l_n	explanation, e_n
strike	The batter did not swing. The ball was in the strike zone.
foul	The batter hit the ball into the stands and it landed in foul territory.
ball	The hitter didn't swing. The ball was outside the strike zone.
none	The video did not load.
out	the batter hit the ball and it was caught by the fielder

Extracted Concepts

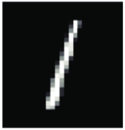
label, l_n	concepts, c_n
strike	{the batter did not swing, the ball was in the strike zone}
foul	{the batter hit the ball, it landed in foul territory}
ball	{the batter did not swing, the ball was outside the strike zone}
out	{the batter hit the ball, it was caught by the fielder}

Text Removed text ○ Completed concepts ▭ Grouped concepts Text Pruned concepts

Consistency, coherence and transfer

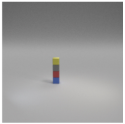
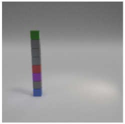


Consistency, coherence and transfer

$\phi_r(\psi($  $),\psi($ $))$

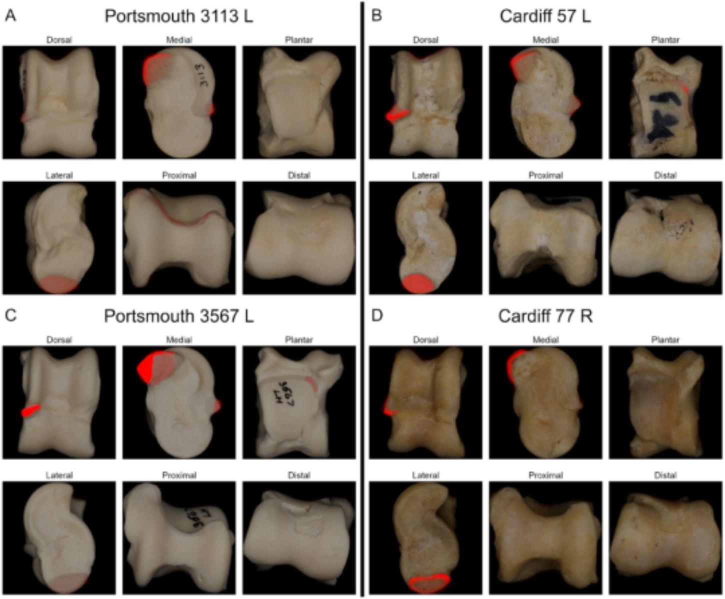
- ψ embeds input to universal space.
- Function ϕ_r approximates relation r.
- Here $r \in \{\text{isGreater}, \text{isEqual}, \text{isLess}, \text{isSuccessor}, \text{isPredecessor}\}$
- Each relation exhibits individual consistency, e.g.
  $\forall X, Y, Z\ :\ \text{isGreater}(X, Y) \wedge \text{isGreater}(Y, Z) \rightarrow \text{isGreater}(X, Y)$
- Can also define consistencies across relations, e.g.
  $\forall X, Y\ :\ \neg\text{isLess}(X, Y) \wedge \neg\text{isEqual}(X, Y) \rightarrow \text{isGreater}(X, Y)$

Consistency, coherence and transfer

$\phi_r(\psi($  $),\psi($  $))$

- ψ embeds input to universal space.
- Function ϕ_r approximates relation r.
- Here $r \in \{\text{isGreater}, \text{isEqual}, \text{isLess}, \text{isSuccessor}, \text{isPredecessor}\}$
- Each relation exhibits individual consistency, e.g.
  $\forall X, Y, Z\ :\ \text{isGreater}(X, Y) \wedge \text{isGreater}(Y, Z) \rightarrow \text{isGreater}(X, Y)$
- Can also define consistencies across relations, e.g.
  $\forall X, Y\ :\ \neg\text{isLess}(X, Y) \wedge \neg\text{isEqual}(X, Y) \rightarrow \text{isGreater}(X, Y)$
- Consistencies preserved even as domain changes.

Example sheep/goat bones



Human-machine saliency investigation