DeepMind

Tell me why! Explanations support learning relational and causal structure

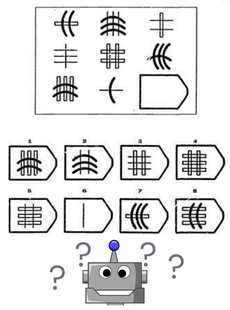
Andrew K Lampinen, Nicholas A Roy, Ishita Dasgupta, Stephanie CY Chan, Allison C Tam,
James L McClelland, Chen Yan, Adam Santoro, Neil C Rabinowitz, Jane X Wang, Felix Hill

RL agents still lack some human skills

While RL agents can accomplish incredible things from reward alone, they still struggle with abstract, causal, and relational tasks that we consider key examples of human intelligence. Why?

Reward is enough for intelligence?

David Silver, Satinder Singh, Doina Precup, Richard S. Sutton



Humans learn from language/explanation

Humans learn from language; in particular explanations, which link concrete situation to abstractions that are:

- Causal
- Generalizable to future situations

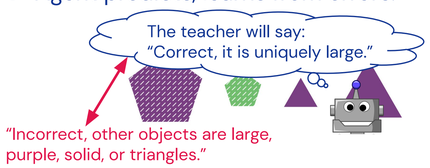
Could explanations help RL agents to learn and generalize?

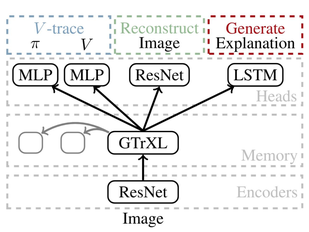
Odd one out tasks

- Choose the object that has a unique feature, among objects that vary along many dimensions. Hard to learn from rewards alone, requires considering all objects + relations!
- Language is not **necessary**, but could explanations help?

Teaching agents to predict explanations

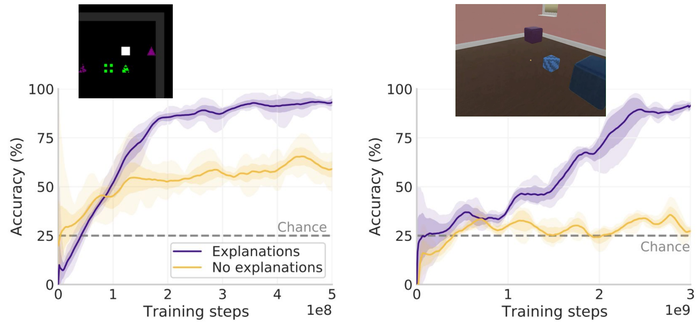
- Teacher (environment) provides explanation responding to agent actions.
- Agent predicts, learns from errors.





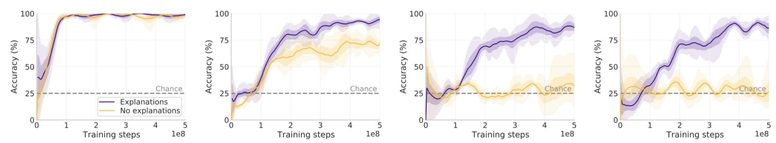
Explanations help learn odd one out tasks

- Agents with explanations (purple) learn tasks well; agents without explanations (yellow) do not.
- In both 2D and 3D environments.



Explanations prevent fixation on shortcuts

- In 2D, agents trained without explanations learn the easy features, but fixate on these inadequate "shortcuts", and fail to learn harder ones.
- Agents trained with explanations learn the full structure.

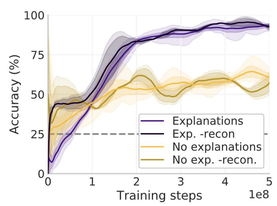


(a) Position. (b) Color. (c) Shape. (d) Texture.

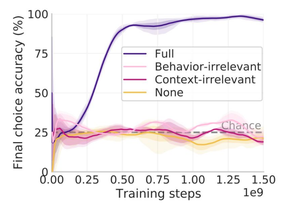
Easier → Harder
For CNNs

Controls, baselines, etc.

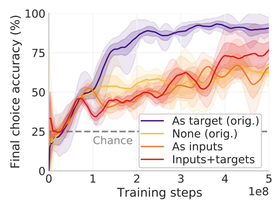
Explanations outperform higher-information unsupervised losses.



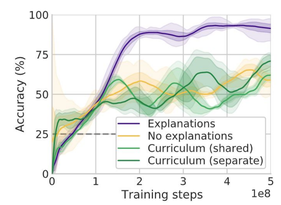
Explanations that respond to agent's behavior are best.



Explanations as input are worse.



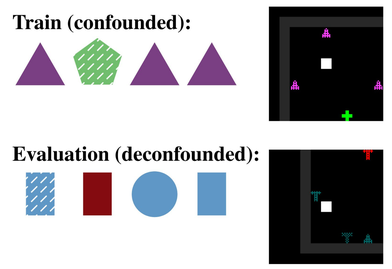
Task curricula are less effective.

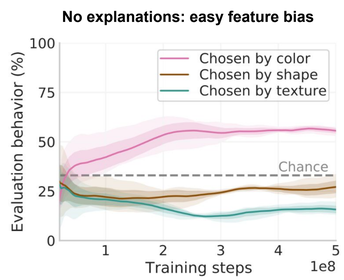


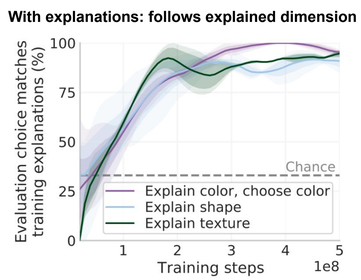
Explanations allow OOD generalization from ambiguous training

What if training is ambiguous?

- Train where a single object is unique along all dimensions.
- Evaluate OOD: a different object unique along each dimension.
- Without explanations, agent primarily uses easy feature.
- With single-dimension explanations, agent generalizes along that dimension.



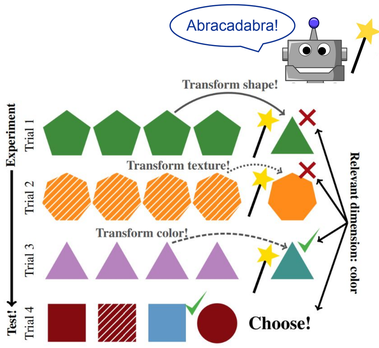


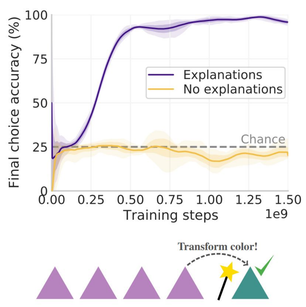


Explanations allow agents to learn to perform causal experiments

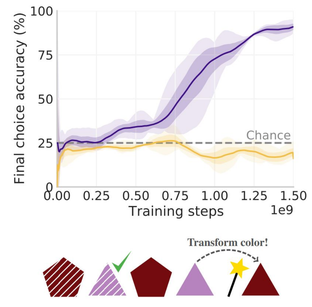
Can agents discover causal structure for themselves?

- Meta-learning setting, where agent experiences four trials.
- It has a magic wand in first three trials that can transform objects.
- Agent has to experiment to discover which feature matters.
- Final trial is a deconfounded test with no wand (and high rewards).
- Explanations enable learning this challenging task.





(b) Easy level structure (bottom) and results (top).



(c) Hard level structure (bottom) and results (top).