DeepMind

Tell me why!

Explanations support learning relational and causal structure

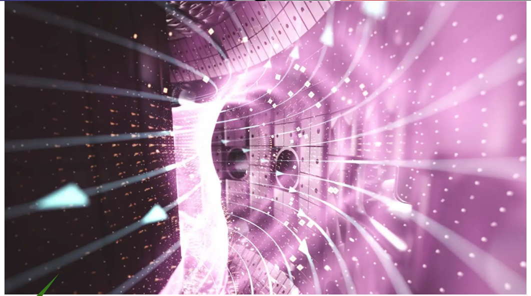
Andrew Lampinen

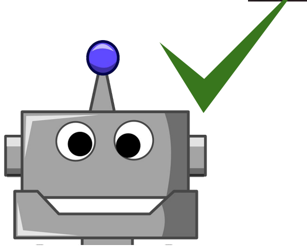
Nick Roy, Ishita Dasgupta, Stephanie Chan, Allison Tam, Chen Yan, Jay McClelland, Adam Santoro, Neil Rabinowitz, Jane Wang, Felix Hill

25/2/2022

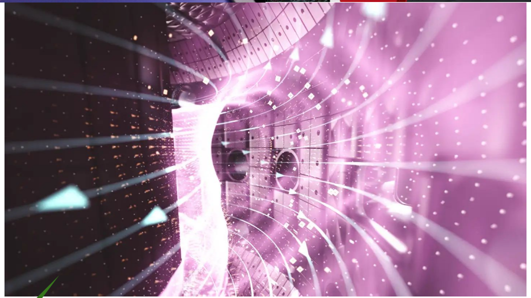
RL: promises and challenges

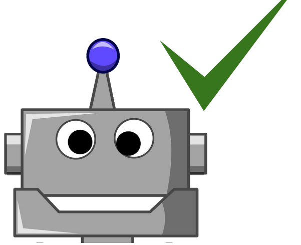
14:32
AlphaStar 177/200 945 758 64 113 940
LiquidTLO 147/172 335 442 61 86 1377
SUPPLY
MINERALS
GAS
WORKERS
ARMY
APM
PRODUCTION





RL: promises and challenges



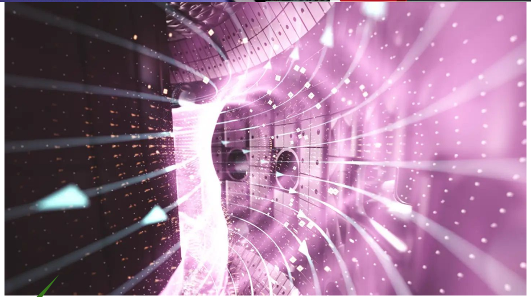


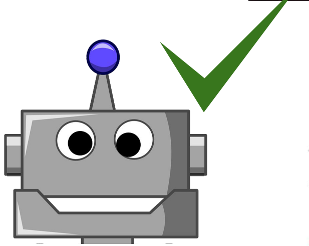
Reward is enough for intelligence?

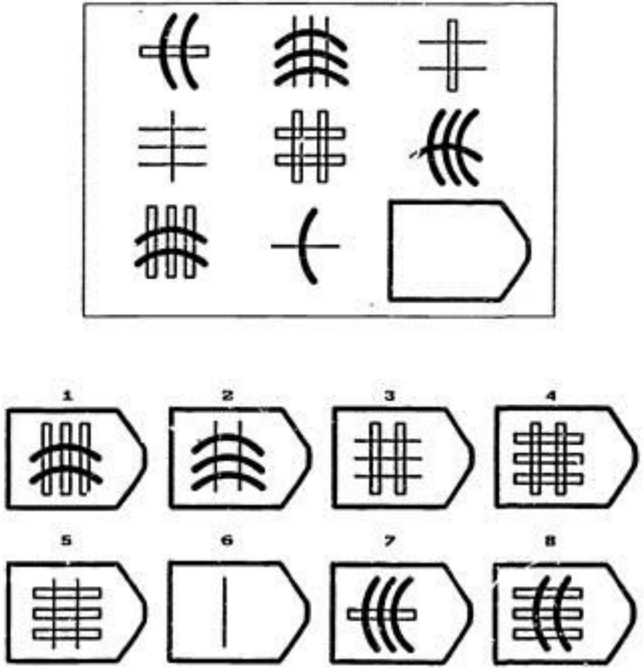
David Silver, Satinder Singh, Doina Precup, Richard S. Sutton

RL: promises and challenges

14:32
AlphaStar 177/200 945 758 64 113 940
LiquidTLO 147/172 335 442 61 86 1377
SUPPLY MINERALS GAS WORKERS ARMY APM PRODUCTION

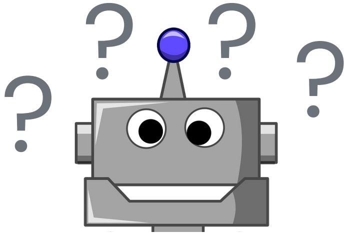






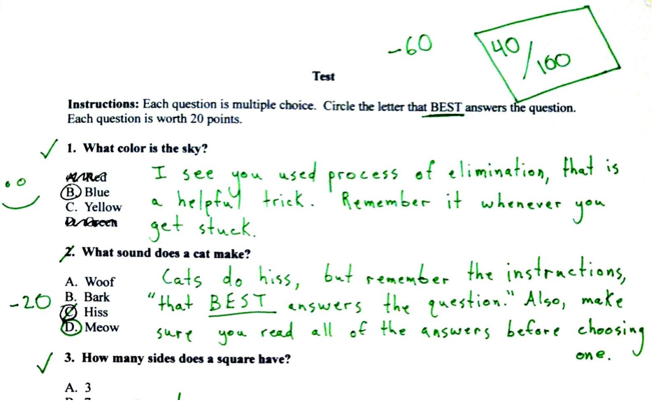
Reward is enough for intelligence?

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



What is different about human learning?

What is different about human learning?



-60

40/100

Test

Instructions: Each question is multiple choice. Circle the letter that BEST answers the question. Each question is worth 20 points.

✓ **1. What color is the sky?**

~~A. Red~~
[✓ B.] Blue
C. Yellow
~~D. Green~~

I see you used process of elimination, that is a helpful trick. Remember it whenever you get stuck.

✗ **2. What sound does a cat make?**

A. Woof
B. Bark
~~C.~~ Hiss
[✓ D.] Meow

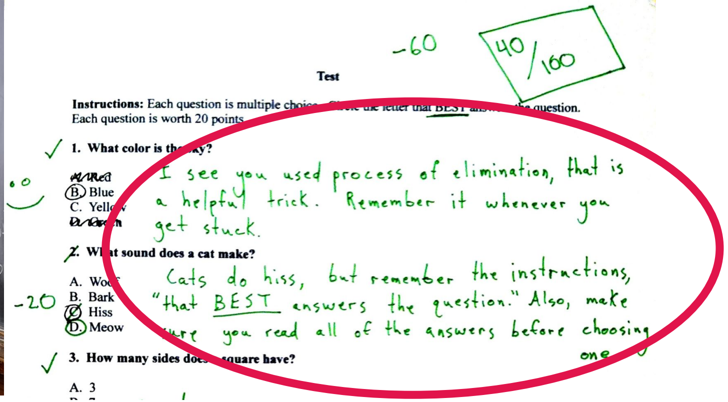
-20

Cats do hiss, but remember the instructions, "that BEST answers the question." Also, make sure you read all of the answers before choosing one.

✓ **3. How many sides does a square have?**

A. 3

What is different about human learning?



-60 40/100

Test

Instructions: Each question is multiple choice. Circle the letter that BEST answers the question.
Each question is worth 20 points.

✓ **1. What color is the sky?**

~~A.~~ Red
[✓ B.] Blue
C. Yellow
~~D.~~ Green

I see you used process of elimination, that is a helpful trick. Remember it whenever you get stuck.

~~2.~~ **What sound does a cat make?**

A. Woof
B. Bark
~~C.~~ Hiss
[✓ D.] Meow

-20

Cats do hiss, but remember the instructions, "that BEST answers the question." Also, make sure you read all of the answers before choosing one.

✓ **3. How many sides does a square have?**

A. 3

Explanation along with reward!

What is an explanation?

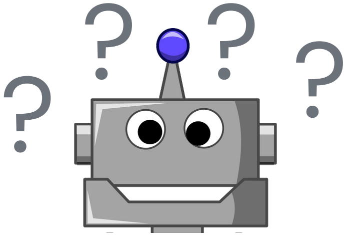
Explanations are intended to *communicate* the links between:

- Concrete situation
- Abstract principles which are:
  - Generalizable
  - Causal

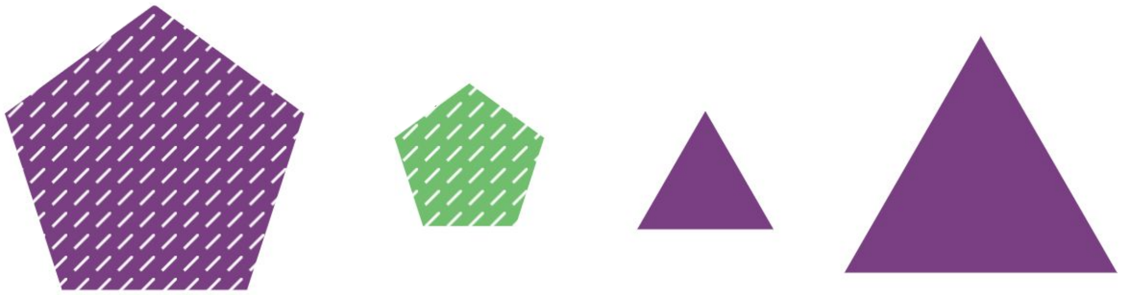
Explanations help us to learn and generalize abstract tasks!

(Lombrozo, 2006; Lombrozo & Carey, 2006; Woodward, 2003)

Could agents learn and generalize better if trained with explanations?

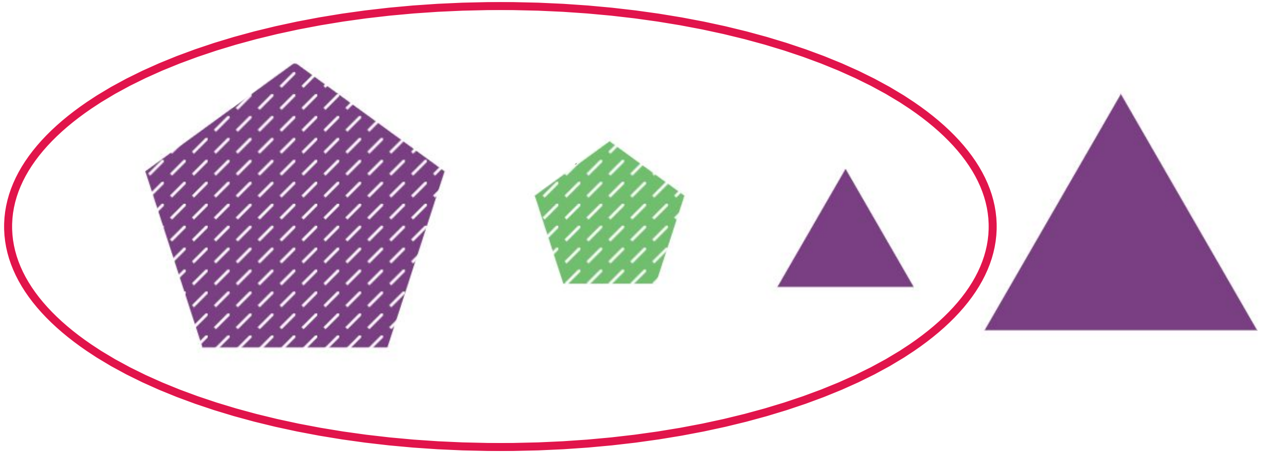


Odd-one-out task: abstraction and relational reasoning



(lots of prior work in cog/neuro, e.g. Stephens & Navarro, 2008; Crutch et al., 2009)

Odd-one-out task: abstraction and relational reasoning

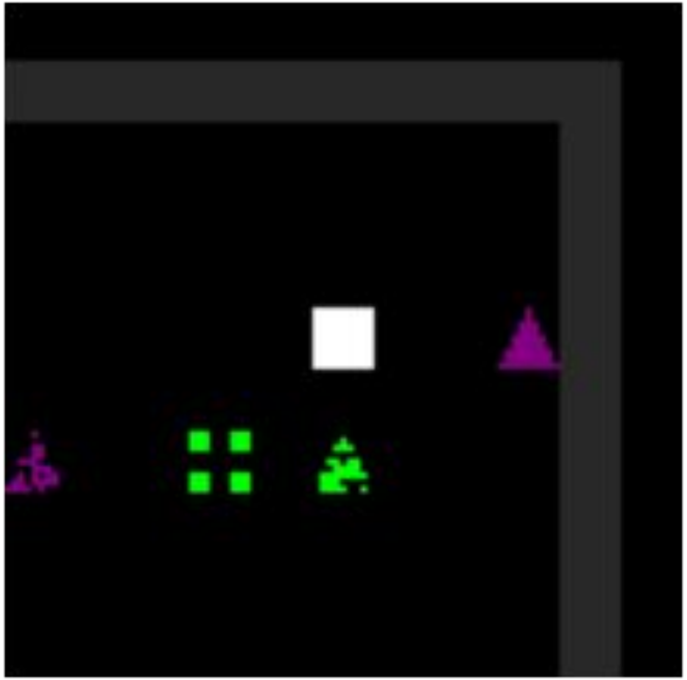


Proper subsets don’t reveal the answer!

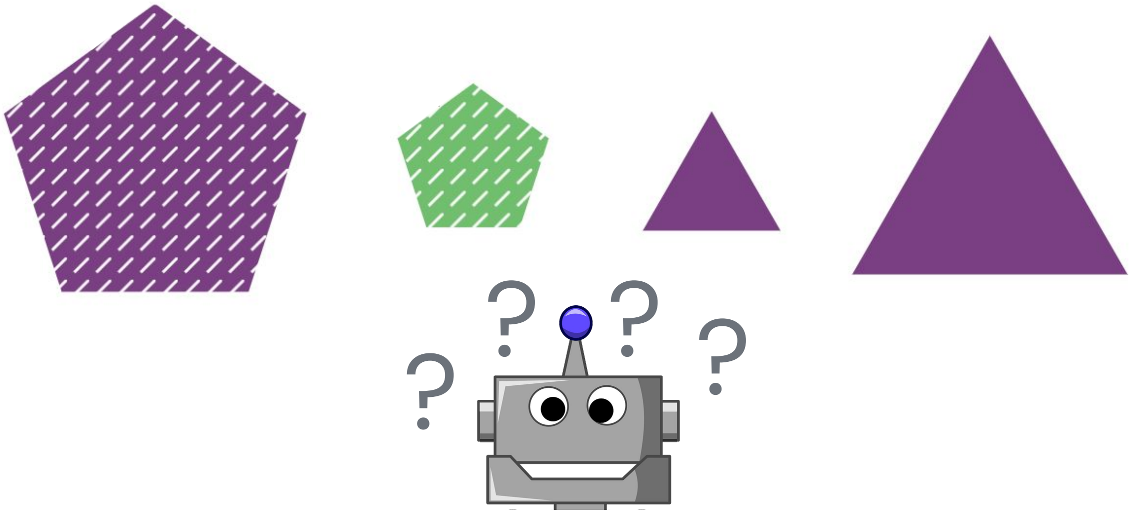
A challenging credit assignment problem from reward alone.

(lots of prior work in cog/neuro, e.g. Stephens & Navarro, 2008; Crutch et al., 2009)

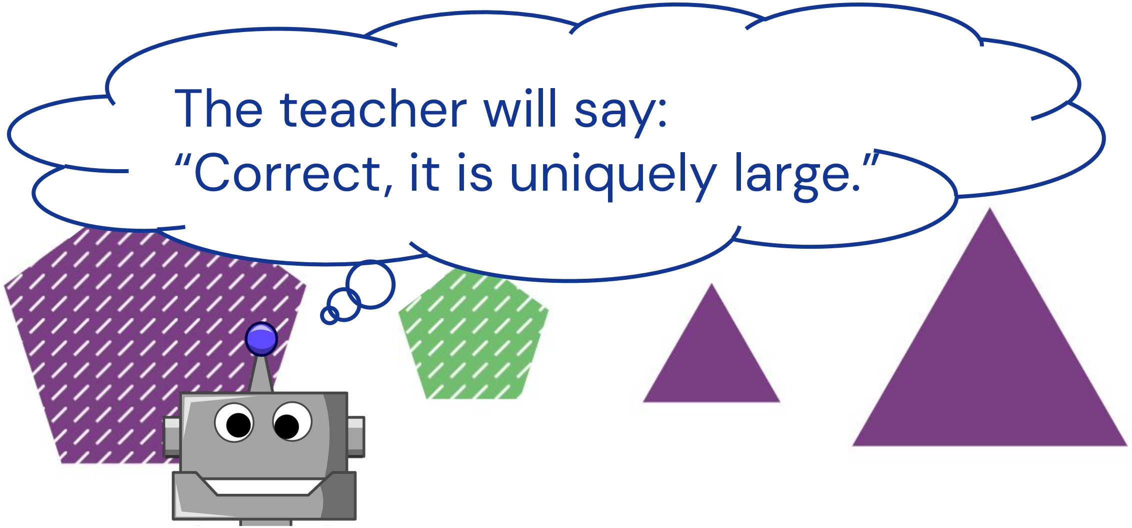
Instantiated in 2D and 3D environments



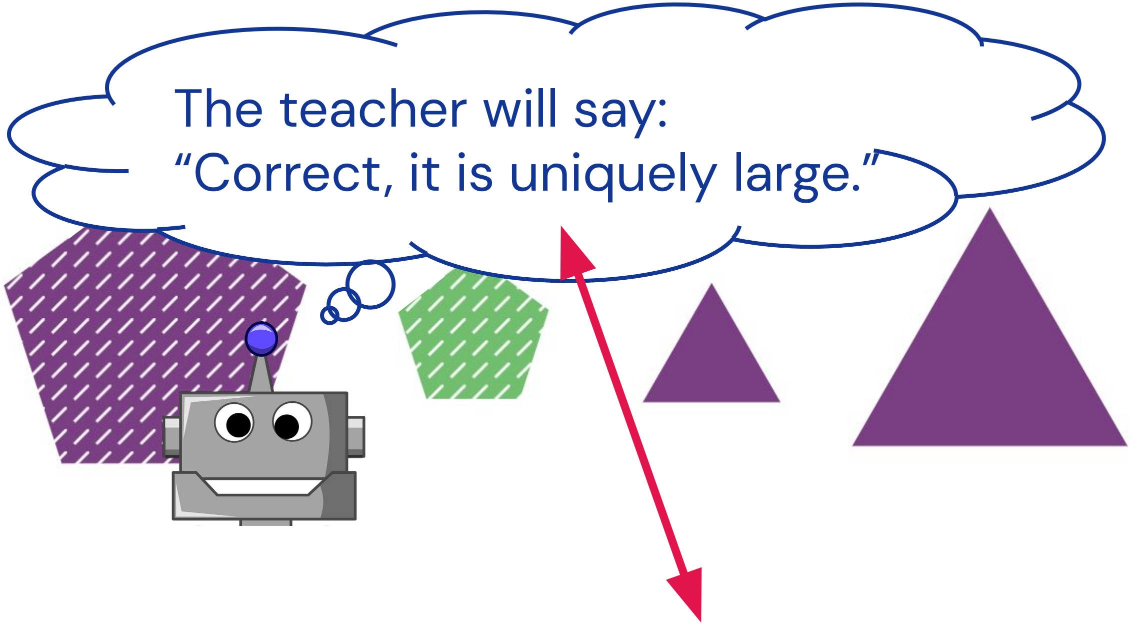
Agents struggle to learn odd-one-out tasks from rewards alone: abstractions and relations are hard to infer!



Predicting explanations during training



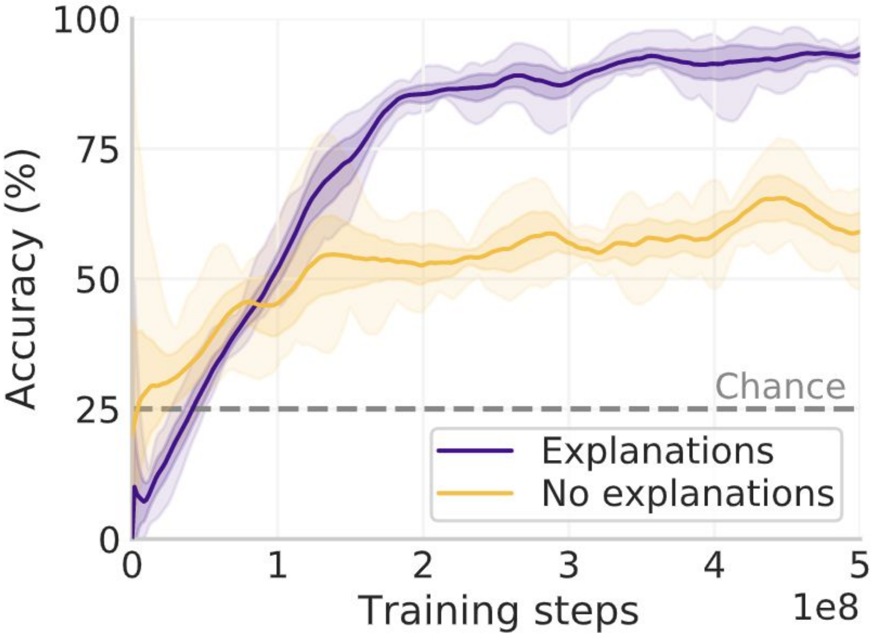
Predicting explanations during training



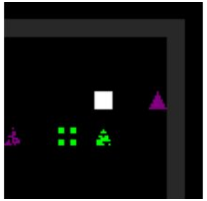
The teacher will say:
"Correct, it is uniquely large."

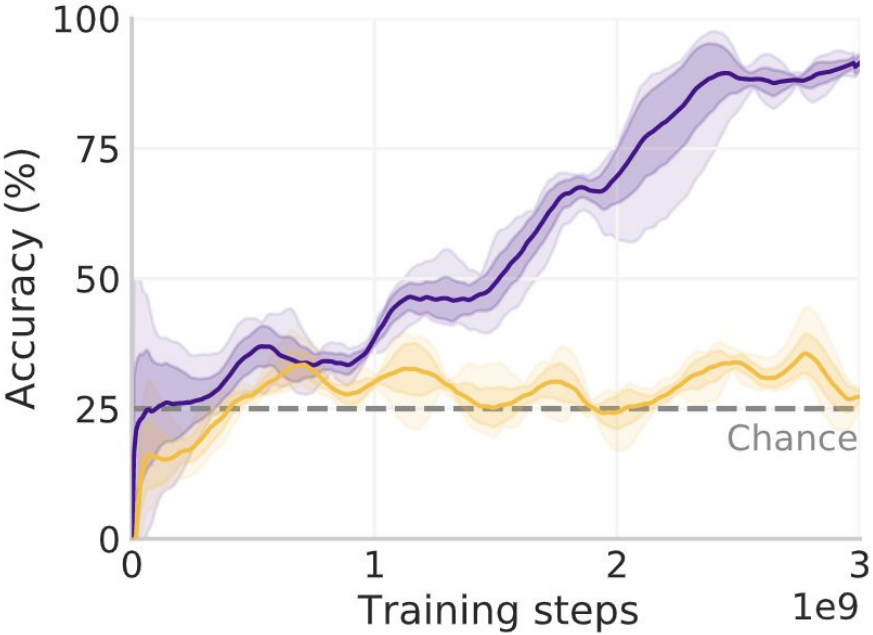
"Incorrect, other objects are large, purple, striped, or pentagons ."

Explanations improve learning of odd-one-out tasks



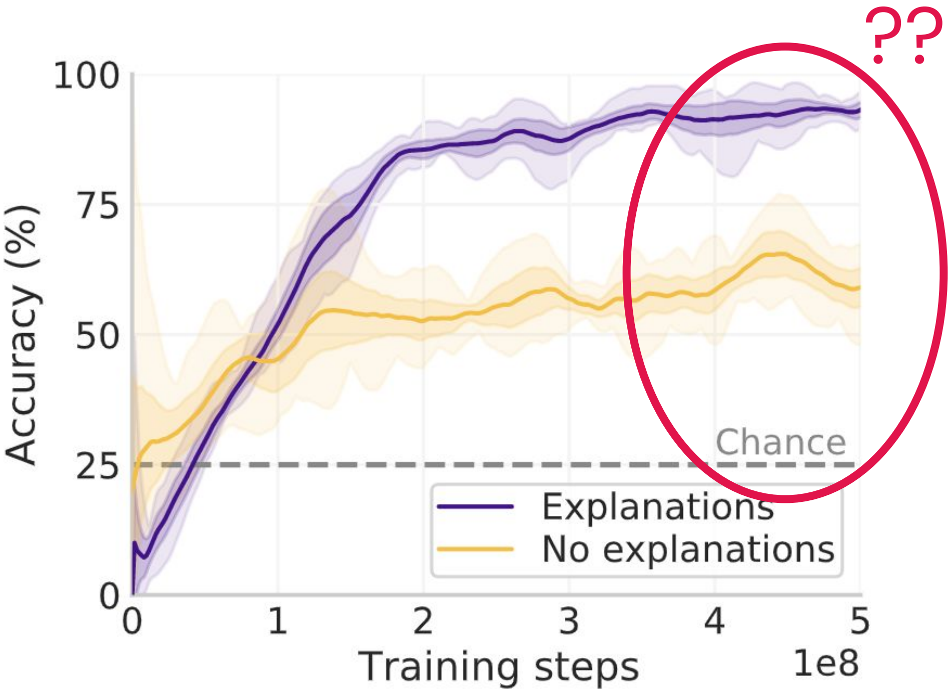
(b) 2D results.



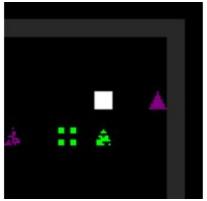


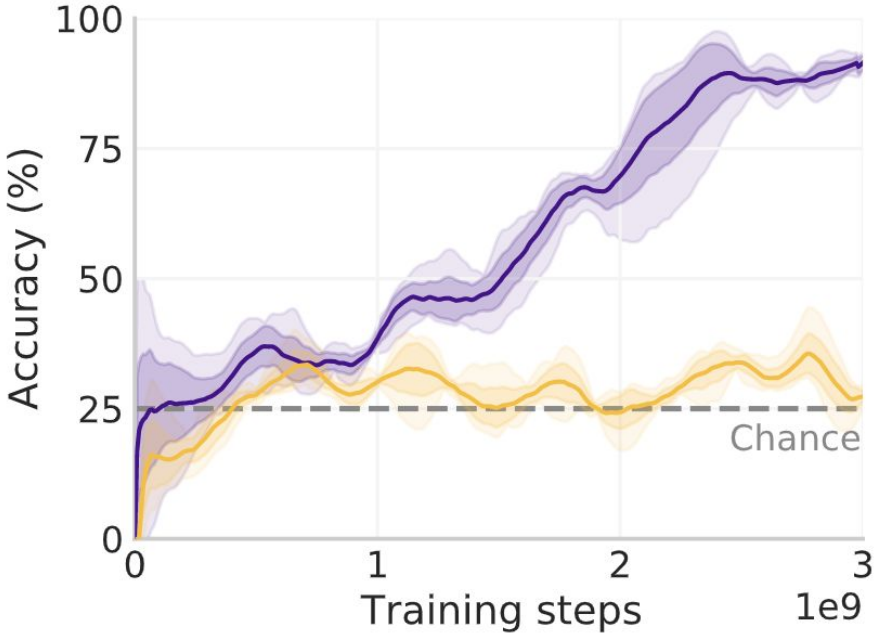
(c) 3D results.

Explanations improve learning of odd-one-out tasks



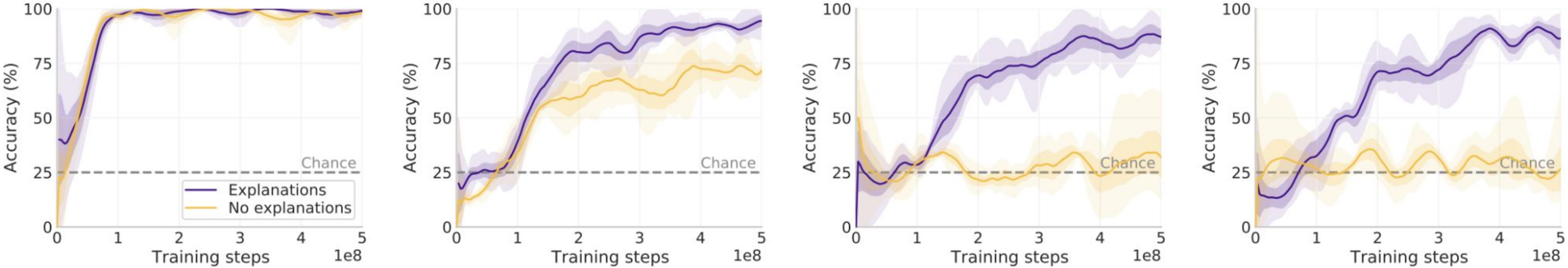
(b) 2D results.





(c) 3D results.

In 2D, explanations help agents move past focus on “shortcut” features to learn more difficult ones



(a) Position.

(b) Color.

(c) Shape.

(d) Texture.

Easier → Harder

For CNNs

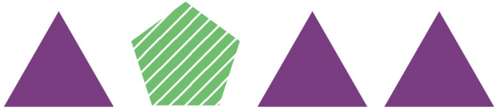
(cf. Hermann & Lampinen, 2020; Geirhos et al., 2020)

Explanations can help with other important challenges (see paper)

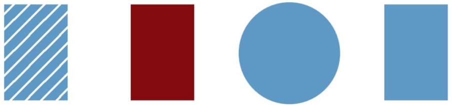
Explanations can help with other important challenges (see paper)

Explanations can overcome ambiguous training to shape how an agent generalizes OOD!

Train (confounded):

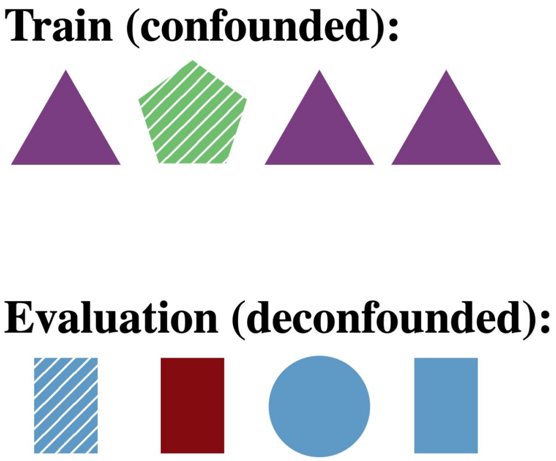


Evaluation (deconfounded):

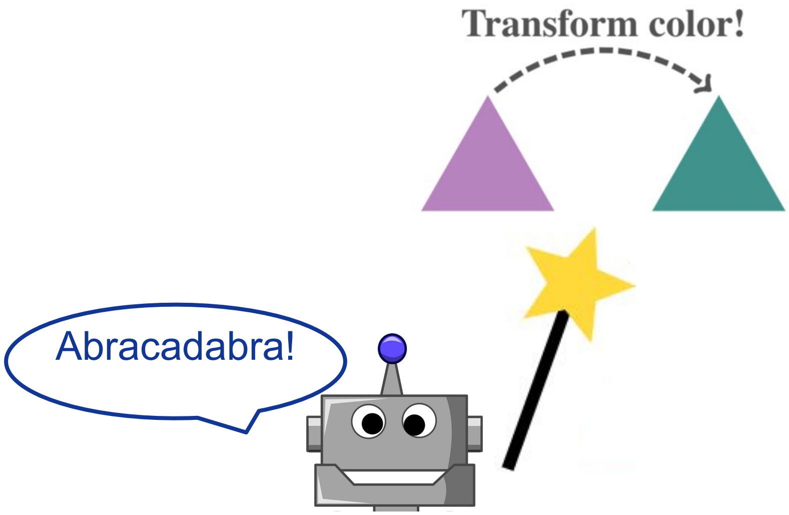


Explanations can help with other important challenges (see paper)

Explanations can overcome ambiguous training to shape how an agent generalizes OOD!

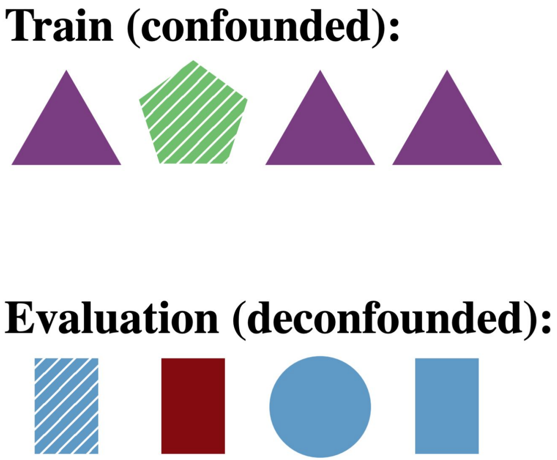


Explanations allow agents to meta-learn how to perform experiments to identify causality!

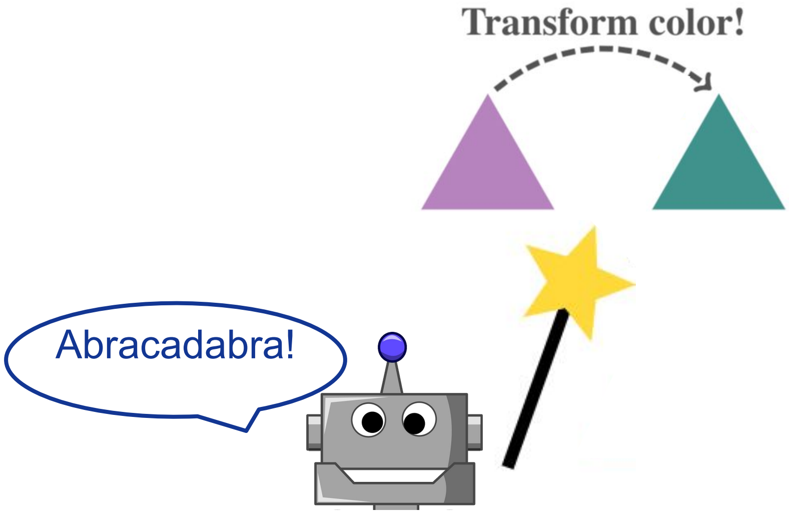


Explanations can help with other important challenges (see paper)

Explanations can overcome ambiguous training to shape how an agent generalizes OOD!



Explanations allow agents to meta-learn how to perform experiments to identify causality!



+ see paper for more analysis, control conditions like unsupervised auxiliary losses, …