Google DeepMind

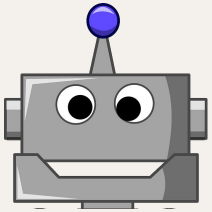
Passive learning of active causal strategies

in agents and language models

Andrew Lampinen

LxMLS
July 19th, 2023

RL agents

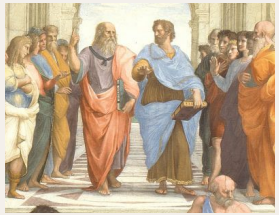


Language Models

Explanations

This work

Philosophy



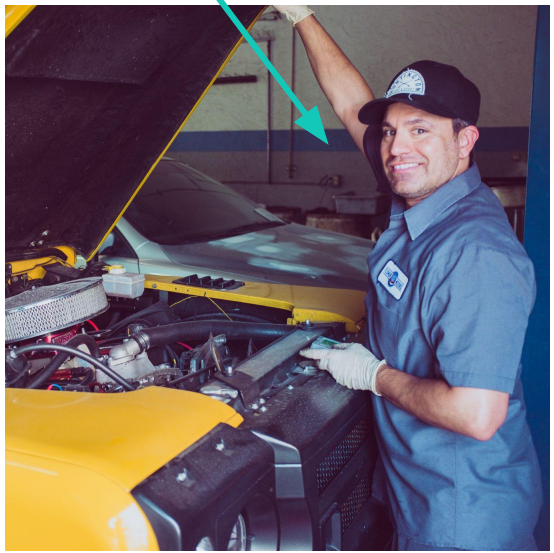
Causality

$$p(Y|do(X=x))$$

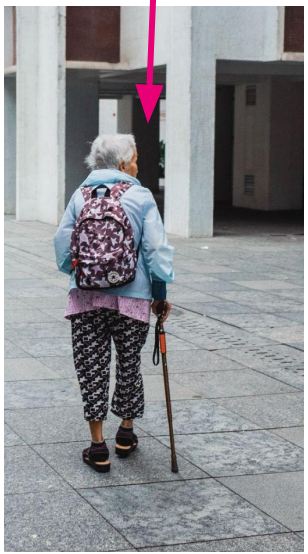
Cognitive science

Passive observation generally can't distinguish causal structure from correlations

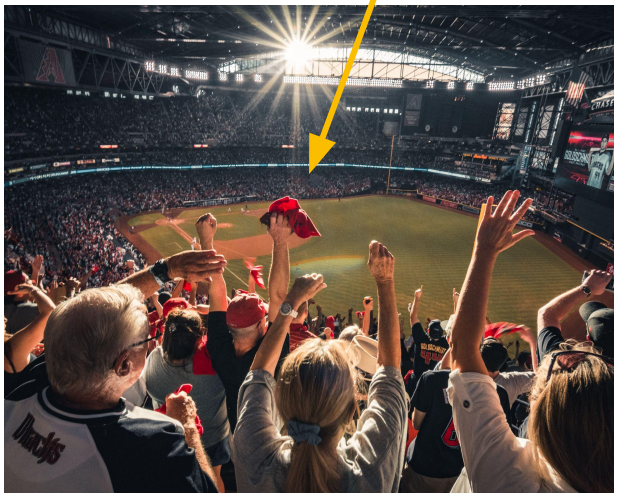
Cars are much more likely to be broken when this guy is around. Kinda suspicious...



People with canes are more likely to have grey hair. Does using a cane make you go grey?

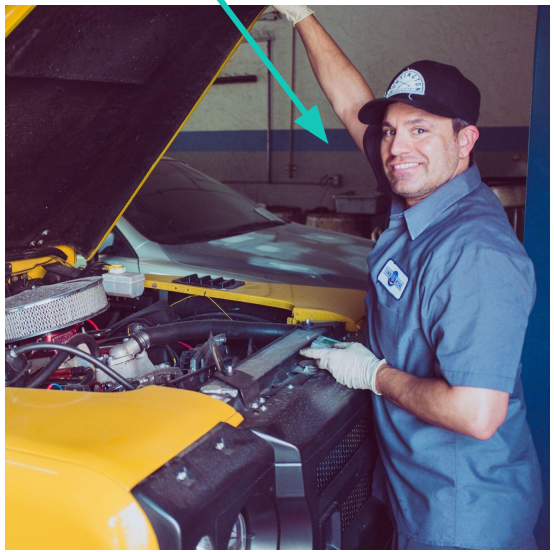


When the team scores, the crowd cheers.

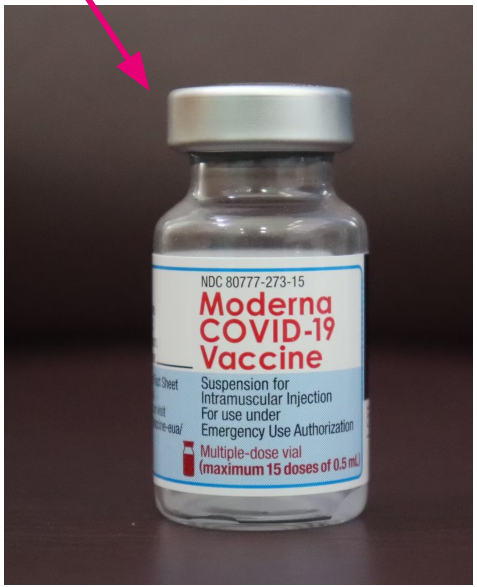


So we need to *do* experiments where we intervene on the world to determine causality!

When I actually brought my car to him, it started working better, rather than breaking!

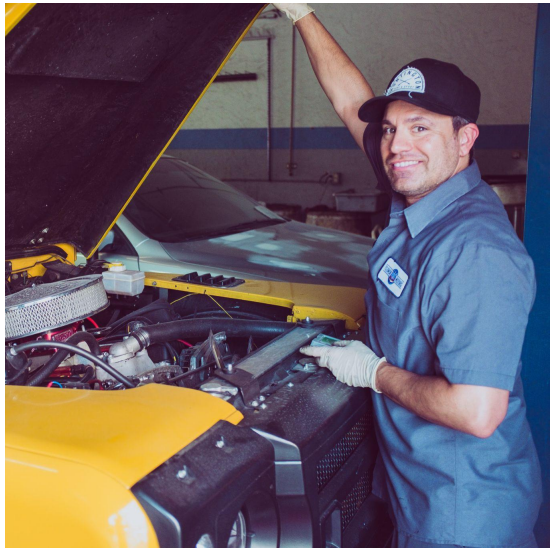


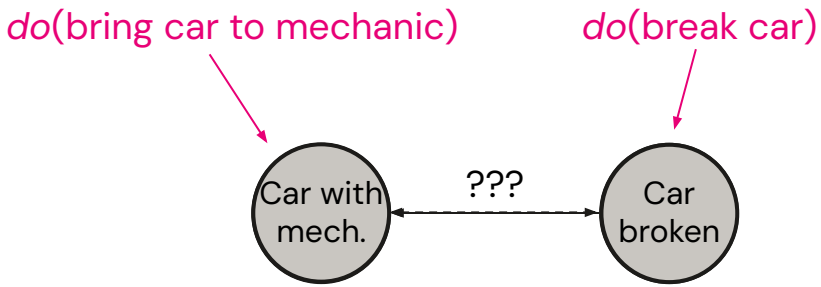
When we give people this vaccine, it lowers their risk of severe covid infections.



More formally, we *do* interventions in causal DAGs

- Causal DAG = Causal Directed Acyclic Graph
  - Nodes = variables = (abstract) states of the world
  - Edges = causal effects (with direction)
- *Do* = operator where you set the state of one (or more) nodes





Pearl's causal hierarchy

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Pearl is pessimistic about modern ML

"This hierarchy, and the formal restrictions it entails, explains why statistics-based machine learning systems are prevented from reasoning about actions, experiments and explanations. [...] the hierarchy denegrades [*sic*] the impressive achievements of deep learning to the level of Association [...] Unfortunately, the theoretical barriers that separate the three layers in the hierarchy tell us that the [...] objective function does not matter. As long as our system optimizes some property of the observed data, however noble or sophisticated, while making no reference to the world outside the data, we are back to level-1 of the hierarchy with all the limitations that this level entails."

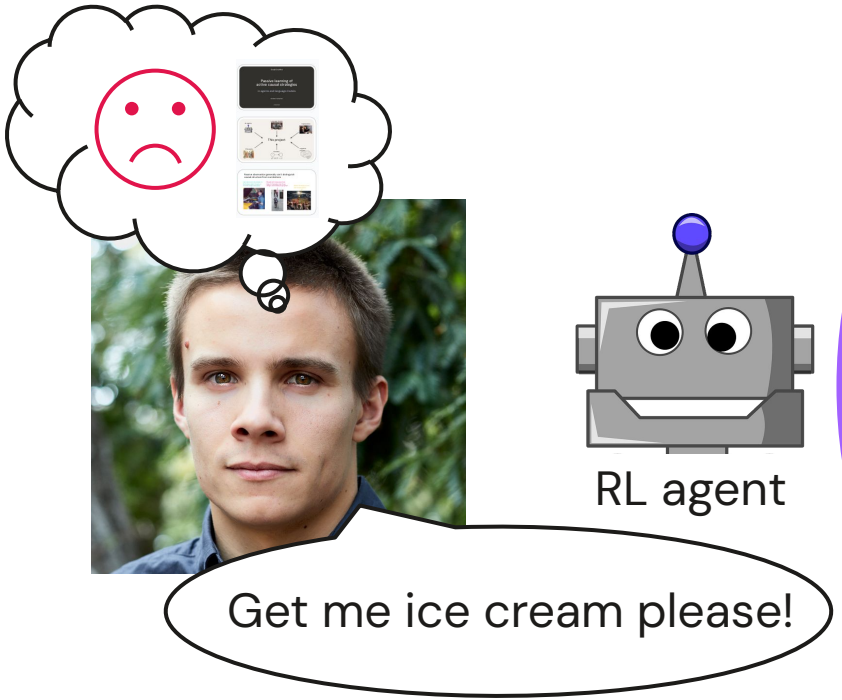
The Three

Level (Symbol)			
1. Associat $P(y\|x)$			
2. Interven $P(y|do(x), $		differently?	Oswald not sh… What if I had not been sm… ing the past 2 years?
3. Counter $P(y_x\|x', y')$			

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

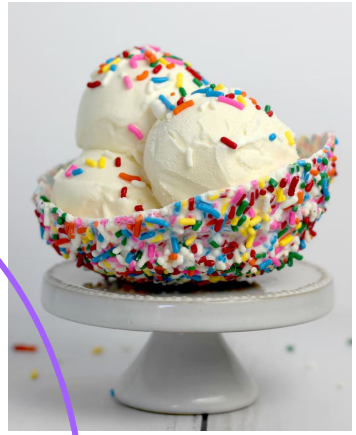
What ML systems *can* intervene?

Reinforcement Learning (RL) agents



RL agent

Live actions
= interventions

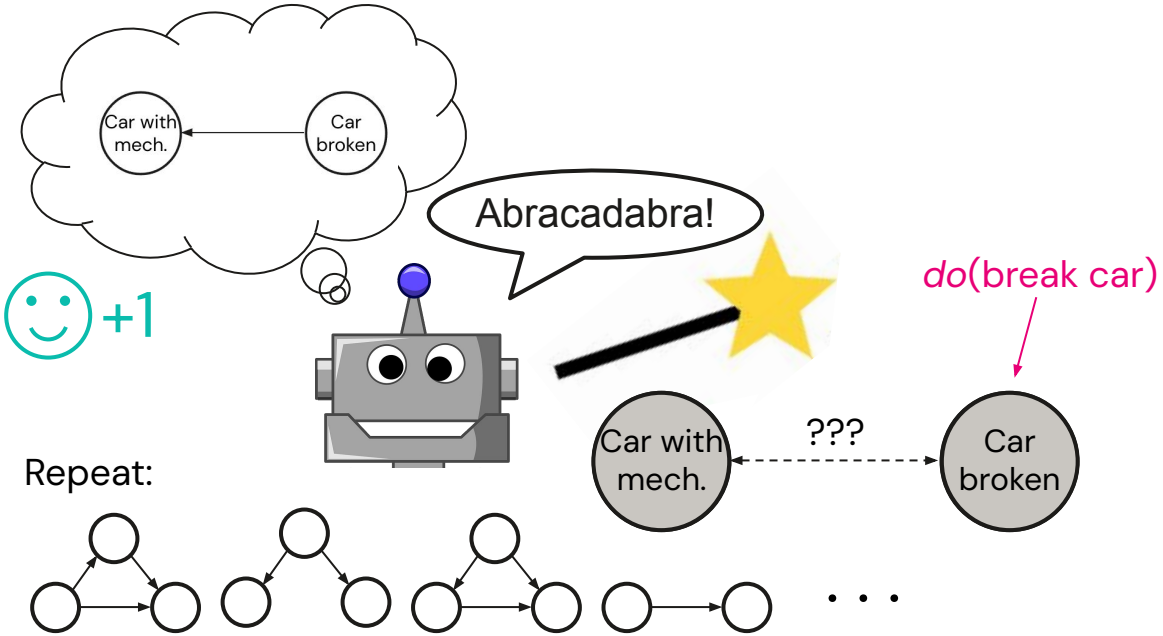


+1

-1

RL to the rescue?

- Because RL agents can *do* things in the environment, they are not fundamentally limited from discovering causality.
- Indeed, lots of prior work shows that RL agents can (meta-)learn how to infer causal structure.



Meta-reinforcement learning of causal strategies

Ishita Dasgupta*[1,4], Zeb Kurth-Nelson[1,2]
Silvia Chiappa[1], Jovana Mitrovic[1], Pedro Ortega[1],
Edward Hughes[1], Matthew Botvinick[1,3], Jane Wang[1]

[1]DeepMind, UK
[2]MPS-UCL Centre for Computational Psychiatry, UCL, UK
[3]Gatsby Computational Neuroscience Unit, UCL, UK
[4]Department of Physics and Center for Brain Science, Harvard University, USA

CAUSAL INDUCTION FROM VISUAL OBSERVATIONS FOR GOAL DIRECTED TASKS

Suraj Nair Yuke Zhu Silvio Savarese Li Fei-Fei
Stanford University

Tell me why! Explanations support learning relational and causal structure

Andrew K. Lampinen [1] Nicholas A. Roy [1] Ishita Dasgupta [1] Stephanie C. Y. Chan [1] Allison C. Tam [1]
James L. McClelland [1] Chen Yan [1] Adam Santoro [1] Neil C. Rabinowitz [1] Jane X. Wang [1] Felix Hill [1]

Learning How to Infer Partial MDPs for In-Context Adaptation and Exploration

Chentian Jiang*
Informatics, University of Edinburgh
Edinburgh, United Kingdom
chentian.jiang@ed.ac.uk

Nan Rosemary Ke
DeepMind
London, United Kingdom
nke@deepmind.com

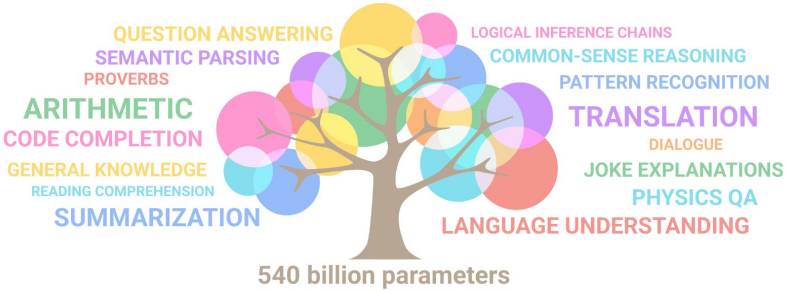
Hado van Hasselt
DeepMind
London, United Kingdom
hado@deepmind.com

It’s not 2019 anymore, nobody cares about RL agents doing toy causal tasks :(

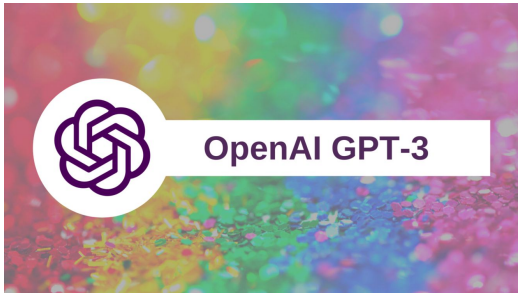
a BigScience initiative

BLOOM

176B params · 59 languages · Open-access



PaLM

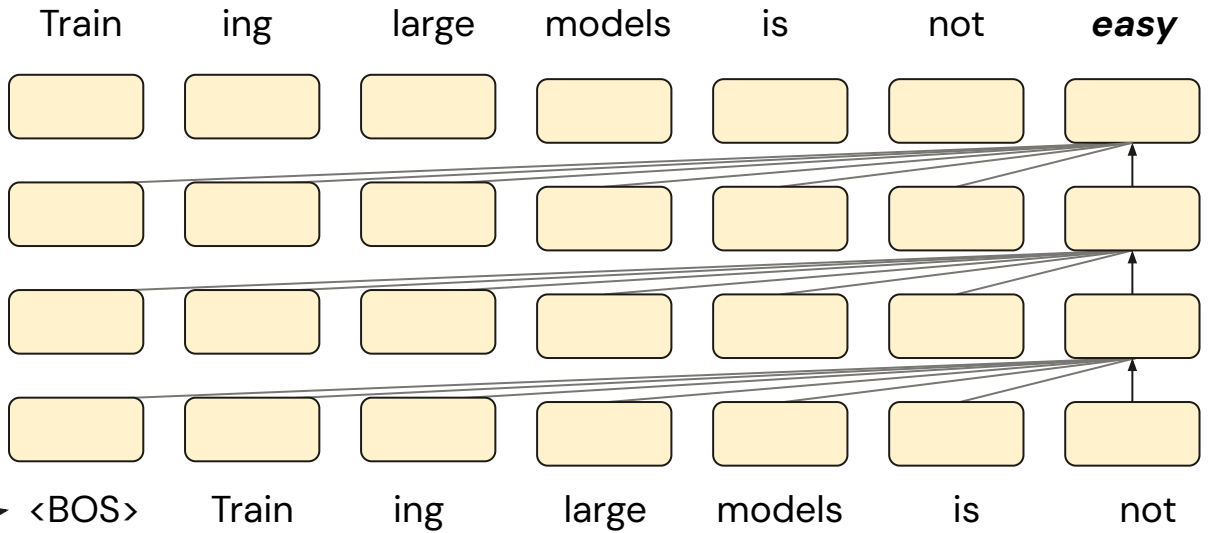


GPT-4

Chinchilla

Language model training: passively predicting the next token in someone else's language

A bunch of text written by some humans



LM training ~= Behavioral Cloning (BC) on human language

(BC is just a fancy way of saying passive imitation of actions that somebody else took)

Language models: fundamentally limited by not intervening?

Gary Marcus ✓
@GaryMarcus

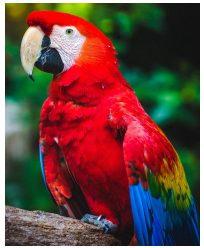
Part of this is right: GPT only knows about correlations within text usage, not about the real world. But scientists –and children as @AlisonGopnik has emphasized--strive to derive causal models of the world, even a world inferred indirectly.

GPT never does that.

> **Connor Leahy** @NPCollapse · Aug 2, 2020
> Replying to @GaryMarcus
> GPT3 can learn to understand its "universe of text", which has correlations to the real world. The same way, humans can't directly learn about the true "quantum physical" world, just approximate it with our macroscopic sense organs.

Judea Pearl ✓ @yudapearl · Aug 5, 2021
3/ trenches of AI research are asking, of course: "What is it?" or "What is the scientific principle by which 'Foundation models' can circumvent the theoretical limitations of data-centric methods as we know them, especially those that hinder generalization across environments?"
3 4 35

Judea Pearl ✓ @yudapearl · Aug 5, 2021
4/ I have tried to pin point the principle, but ended up in failure; the limitations as we know them are not mentioned in the papers I examined, nor are concepts such as "causation" "Data Fusion" "Transportability" "data-centric" etc. Can some readers enlighten us?
4 3 50

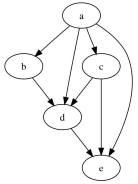


Causal Parrots: Large Language Models May Talk Causality But Are Not Causal

MORITZ WILLIG* and MATEJ ZEČEVIĆ*, Technical University of Darmstadt, Germany
DEVENDRA SINGH DHAMI, Technical University of Darmstadt and hessian.AI, Germany
KRISTIAN KERSTING, Technical University of Darmstadt, hessian.AI and DFKI, Germany

But LMs* can sometimes do sophisticated, interactive things that one might think** require causal understanding!

- LMs* provide useful priors for causal reasoning mechanisms, e.g. for identifying causal structures from data.
- LMs can be prompted to interactively use tools (e.g. APIs) to achieve a task.
- **Are they really doing generalizable causal reasoning, or just parroting causal structures observed in training?
- *Many of these models were trained/tuned with tool interactions and/or RL, and/or ??
  - Is interactive training unlocking these causal abilities?



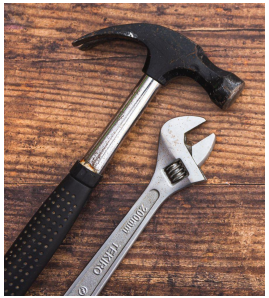
Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

Emre Kıcıman*
Microsoft Research
emrek@microsoft.com

Robert Ness
Microsoft Research
robertness@microsoft.com

Amit Sharma
Microsoft Research
amshar@microsoft.com

Chenhao Tan
University of Chicago
chenhao@uchicago.edu



Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick Jane Dwivedi-Yu Roberto Dessì† Roberta Raileanu
Maria Lomeli Luke Zettlemoyer Nicola Cancedda Thomas Scialom

Meta AI Research †Universitat Pompeu Fabra

Chat Plugins Beta

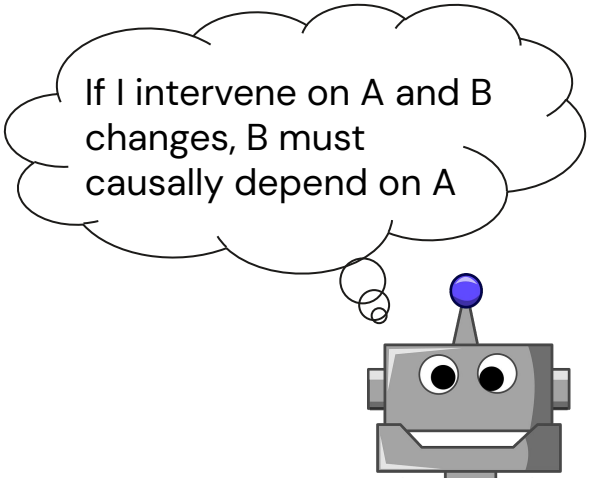
Learn how to build a plugin that allows ChatGPT to intelligently call your API.

LMs are (mostly) passively trained.

Why do they show some behaviors that seem causal?

Two routes to causal understanding from passive learning

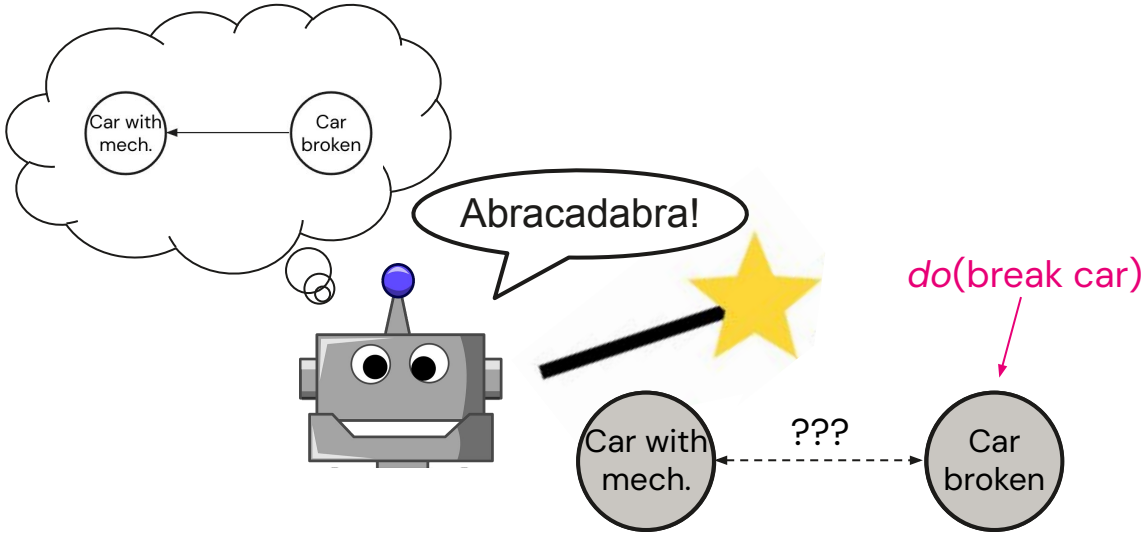
#1: Generalizable causal strategies



#2: Explanations

Idea #1: Higher-level causal strategies may be learnable from observing what someone else has done

- Agents certainly need to *do* interventions **at test time** to be sure of inferring the test-time causal structures, and to take advantage of that knowledge.
- However, it's not clear whether they need to do so **at train time**.
- Passive data can still be interventional (e.g. recordings or descriptions of someone else's experiments).
- Could the agents learn, just from BC, a **strategy** for experimenting + exploiting that would generalize?



Is it possible to learn & generalize causal strategies from passive BC on expert data?

Idea #2: Explanations can highlight causal structure

Explanations are intended to *communicate* the links between:

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

Explanations are designed to highlight causal structure to help us learn!

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

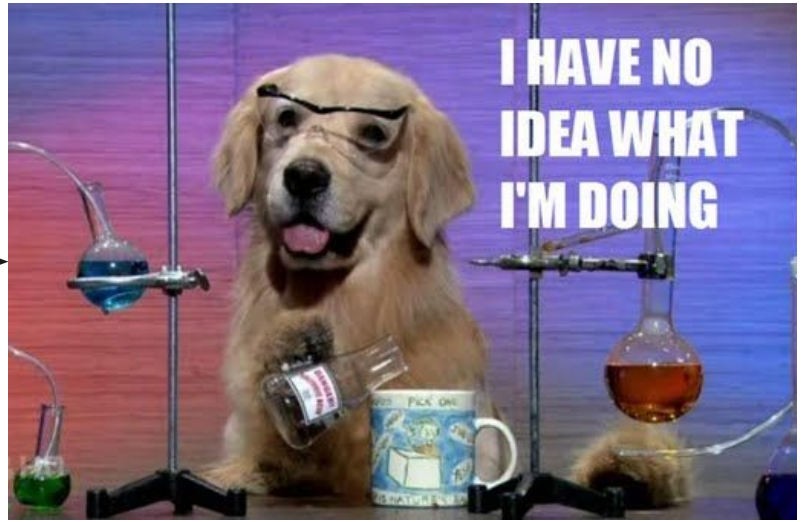
Can explanations support causal learning?

Metaphorically, can you learn to be a scientist just by reading enough books explaining experiments?

Passive learning

?

Active sciencing and new discoveries!

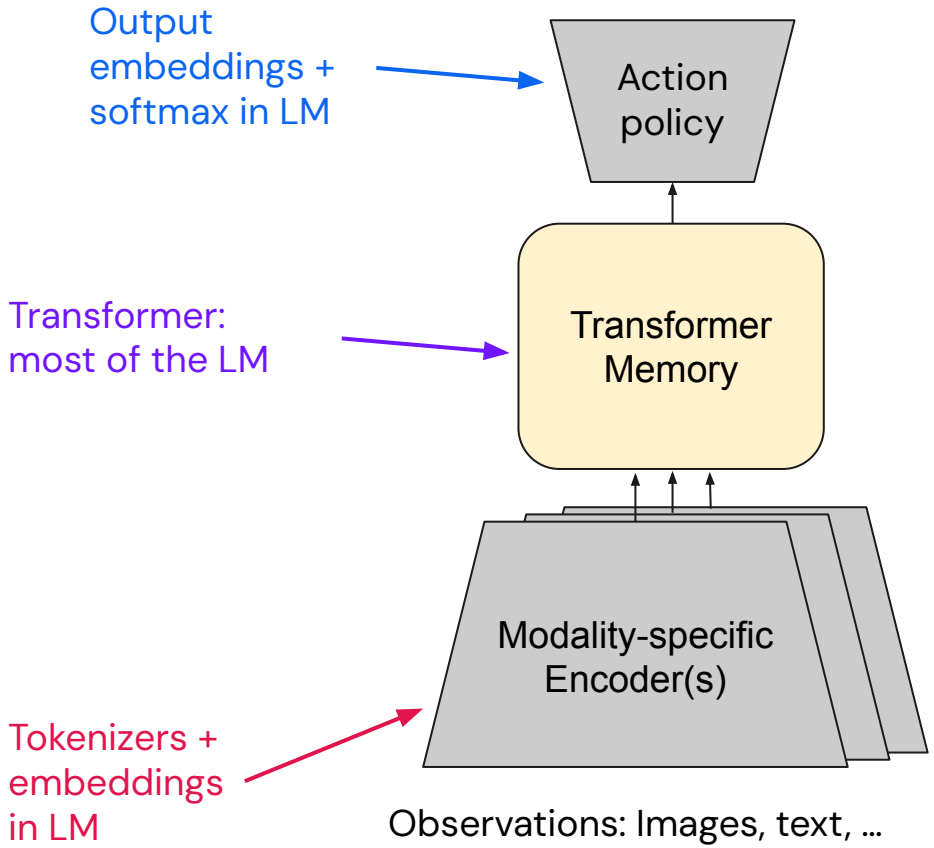


Google DeepMind

1 A simple causal DAG environment

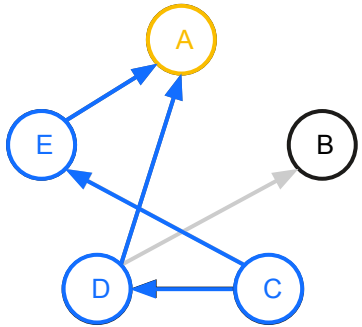
A LM-like agent architecture

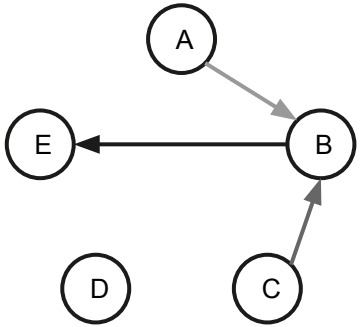
- Test a fairly standard agent architecture (for RL), with input encoding, a memory, and a policy.
- Use a Transformer for the memory, as in many recent RL papers.
- This is actually not so different from the architecture of a language model, except that the ratio of parameters in the memory vs. encoders is higher in LMs (especially large/deep ones).

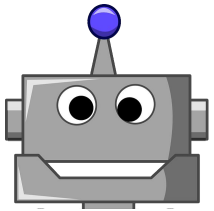


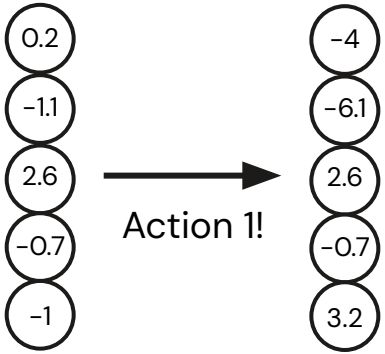
A simple, clean test environment

- Environment consists of an underlying causal DAG over 5 variables {A, B, C, D, E}.
- Variable values are set by linear effect of ancestors (edge weight * ancestor value) + noise, followed by a nonlinearity.
- Agent actions correspond to interventions: set one variable to a large positive or negative value. Agent observes the initial values and outcomes.
- Graph is resampled every episode.
- Episodes consist of a series of trials in two phases:
  - Experimentation phase: the agent is allowed to perform interventions and see the outcomes, without any immediate goal, to (implicitly) infer graph structure. (5 trials)
  - Exploitation phase: agent is given a goal variable to maximize, and then is rewarded with the value of that variable after each intervention. (2 trials)





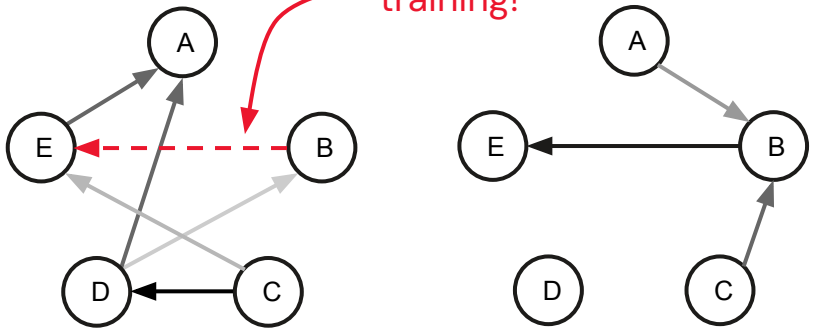




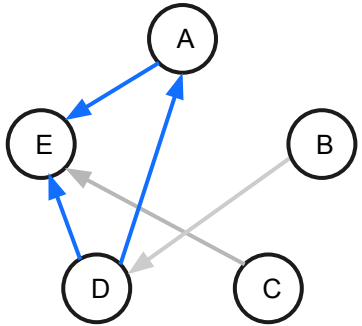
Train with BC; train/test split by causal dependencies

- Train via BC on data from expert that intervenes on each variable once during experimentation, then acts optimally during exploitation.
- In training data, make sure that node D is *never* an ancestor (directly or indirectly) of node E.
- Test on maximizing E in situations where D *is* an important ancestor of E (i.e. high total edge weight from D to E); either where:
  - **Eval. Target:** D is the optimal node to intervene on to maximize E.
  - **Eval. Path:** D is on the key path, but one of its ancestors may be a better intervention.
- This is challenging! Agent never sees **any** situation in training where D has **any** effect on E. Can it generalize to these situations at test?

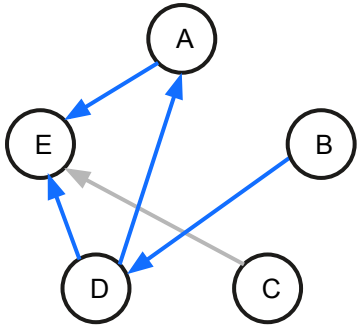
Train:



Eval. Target:

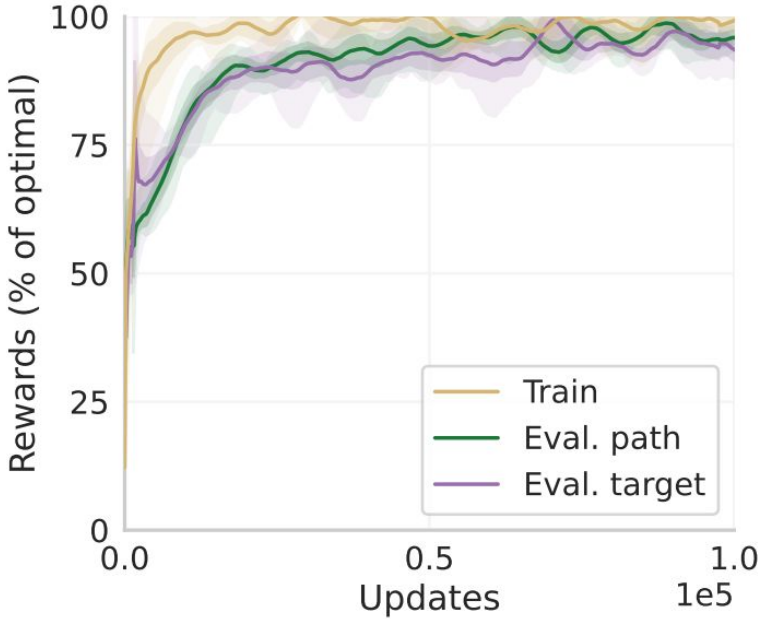


Eval. Path:



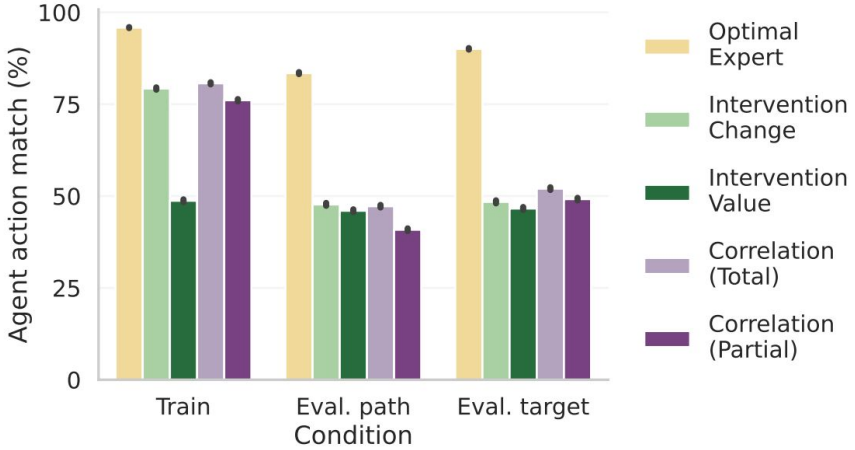
Basic results:

- Over the course of passive training, we run active evaluations.
- Agents quickly learn to achieve near-optimal rewards during the exploitation phase in training and both evaluation conditions.



How are agents doing it?

- Simpler approaches (rather than inferring and reasoning over the causal DAG) could achieve some exploitation performance:
  - Heuristics:
    - Remembering values of nodes after intervention; then repeat the action that achieved the highest outcome value during experimentation on the node that's now the goal.
    - Use the intervention that yielded largest *change* in the goal node; reverse if change was negative.
  - Correlational statistics:
    - Fit regressions from all nodes individually to target node; choose the largest effect for an intervention (total correlation).
    - As above, but control for the effect of other nodes (partial correlation).
- Agent matches optimal causal strategy much better than any of these simpler baselines.

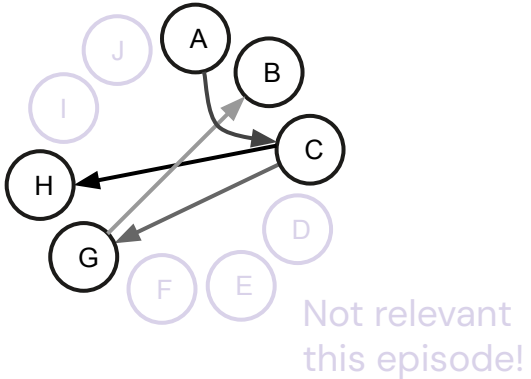


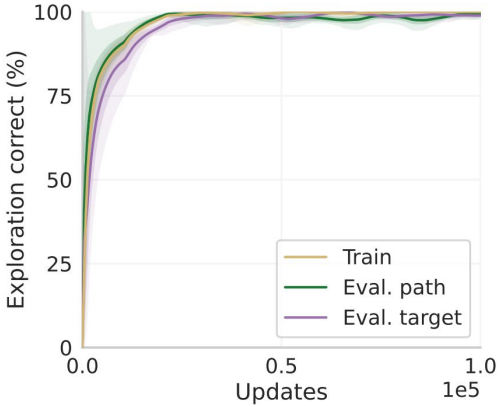
Agents can passively learn active causal experimentation + exploitation strategies!

… at least in these toy environments

Adaptive experiments

- In the above experiments, the expert tried every possible intervention; that's not very scalable.
- Human scientists rely on prior domain knowledge to constrain our hypothesis space.
- We tried a simple version of this: add a bunch of extra variables; each episode, give a multi-hot cue for which variables are relevant (=in DAG) and have the expert only experiment on those.
- However, all variables (relevant and irrelevant) are still included in the agent observations + can be intervened on.
- Hold out subsets of variables (as well as particular causal dependencies as before).
- Can agent generalize to experiment correctly on only the relevant variables in a novel subset at test time, and thereby discover and exploit a dependency that's never appeared in the data?
- Yes, agent rapidly learns to generalize cued exploration.
    - (Also still performs well on exploitation phase, see the paper.)





Google DeepMind

2 Odd-one-out interventions

(

These environments and concepts were introduced in some of our prior work with RL agents

---

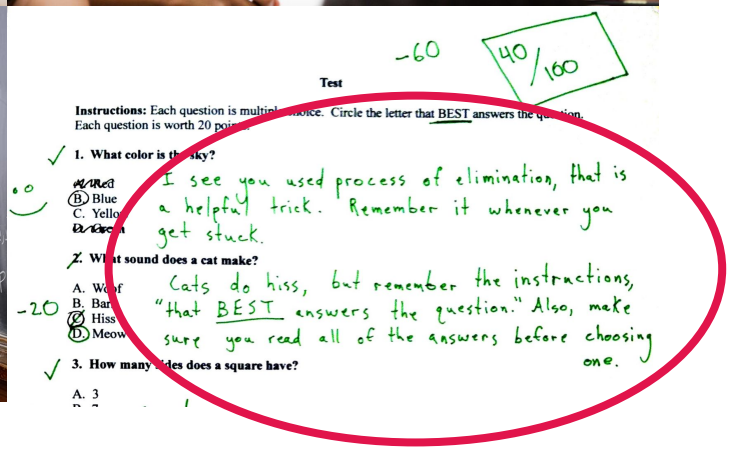
Tell me why! Explanations support learning relational and causal structure

---

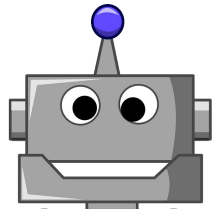
Andrew K. Lampinen [1] **Nicholas A. Roy** [1] **Ishita Dasgupta** [1] **Stephanie C. Y. Chan** [1] **Allison C. Tam** [1]
James L. McClelland [1] **Chen Yan** [1] **Adam Santoro** [1] **Neil C. Rabinowitz** [1] **Jane X. Wang** [1] **Felix Hill** [1]

… so now will have a brief interlude to introduce them

Human learning is pedagogical, and focuses on explanation

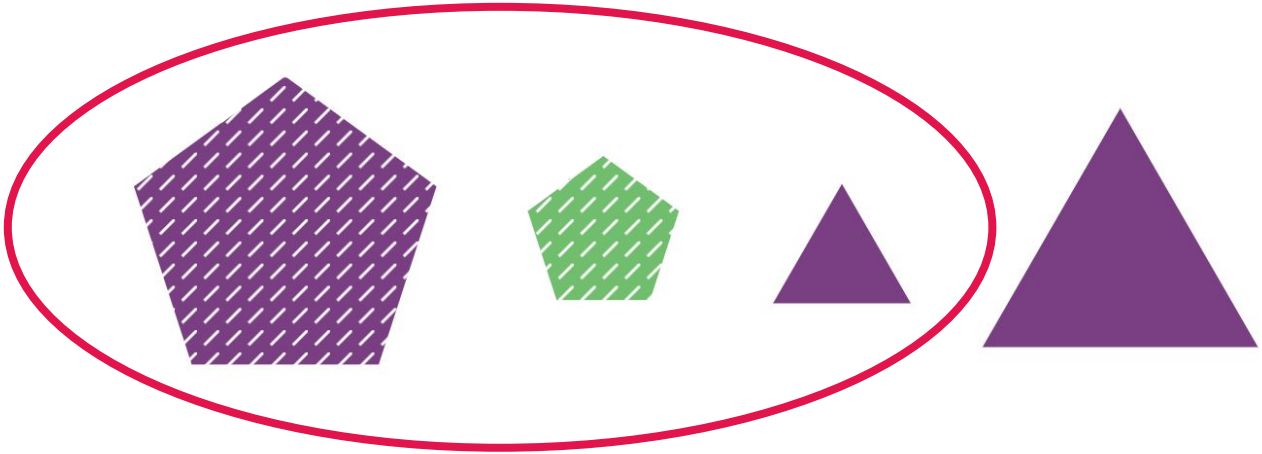


+1



-1

Odd-one-out tasks: abstraction & relations

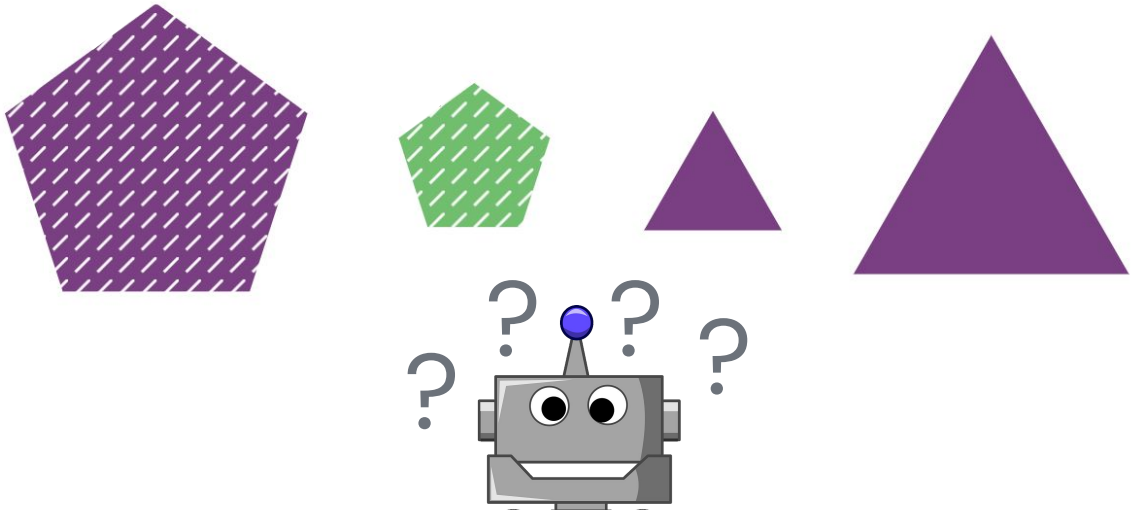


Proper subsets don’t reveal the answer!

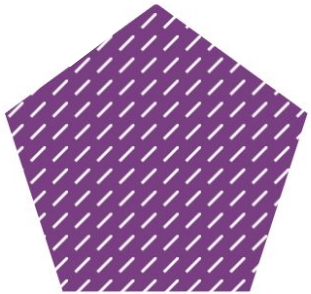
A challenging credit assignment problem from reward alone.

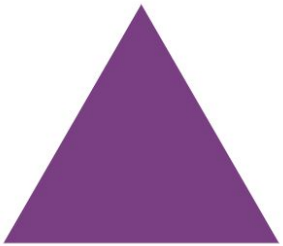
(lots of prior work in cog/neuro, e.g. Stephens & Navarro, 2008; Crutch et al., 2009)

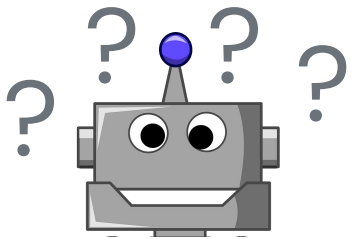
RL agents struggle to learn these tasks from rewards alone!

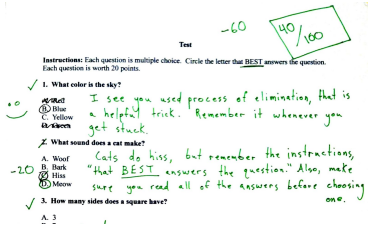


Explanations as auxiliary signals



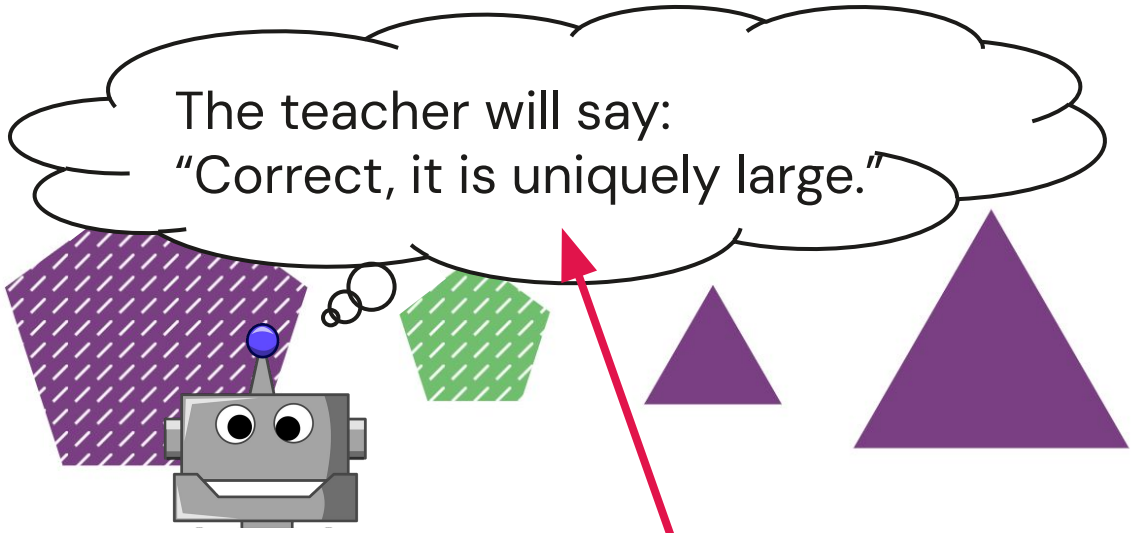






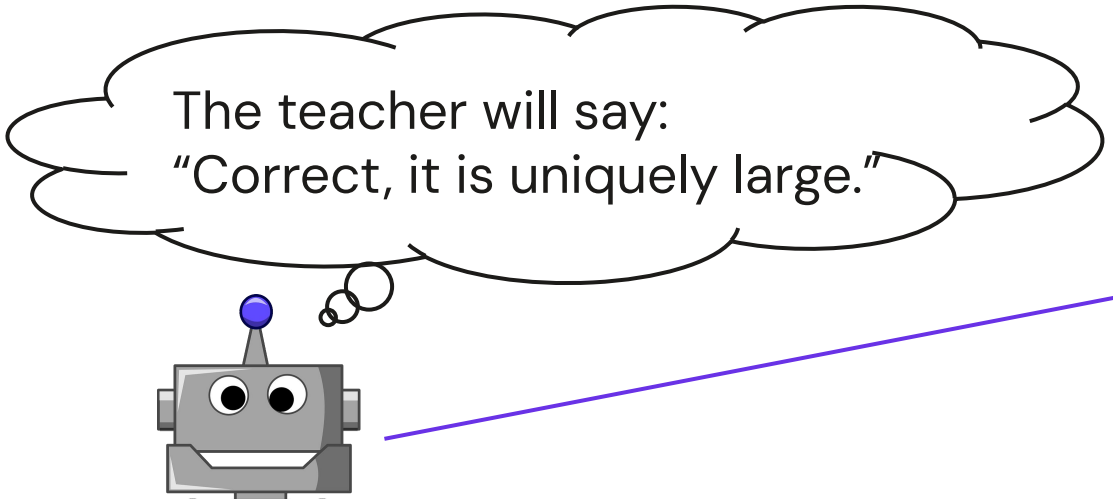
*Are descriptions explanations?
Arguable, but ours at least pragmatically focus on generalizable causal features.

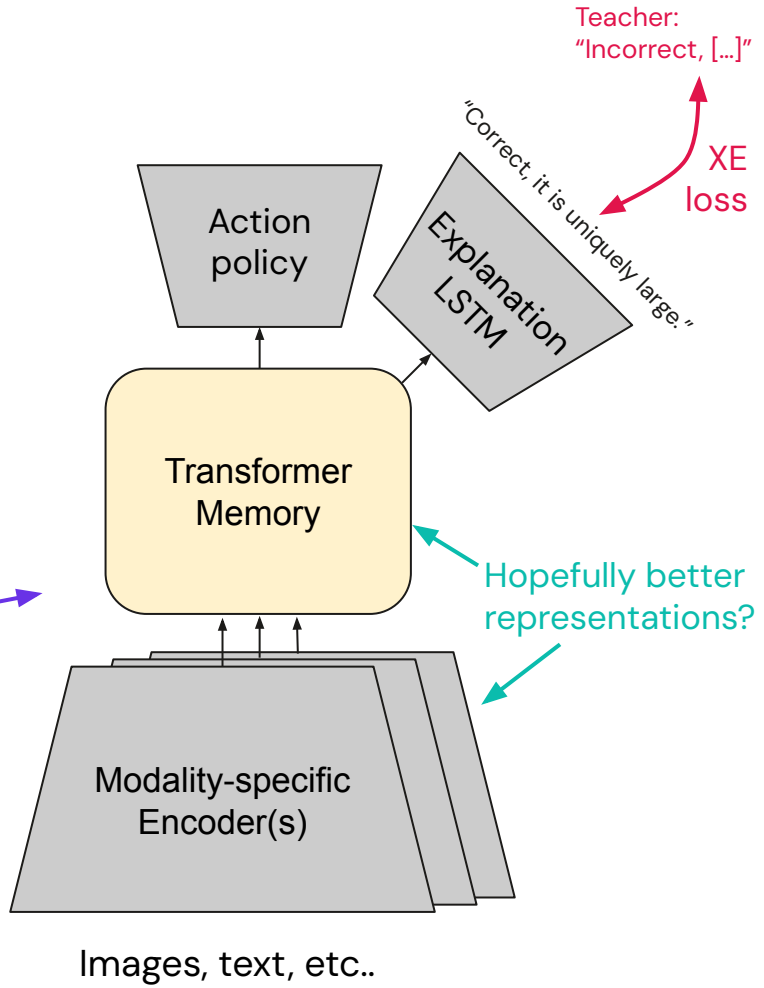
Predicting explanations during training



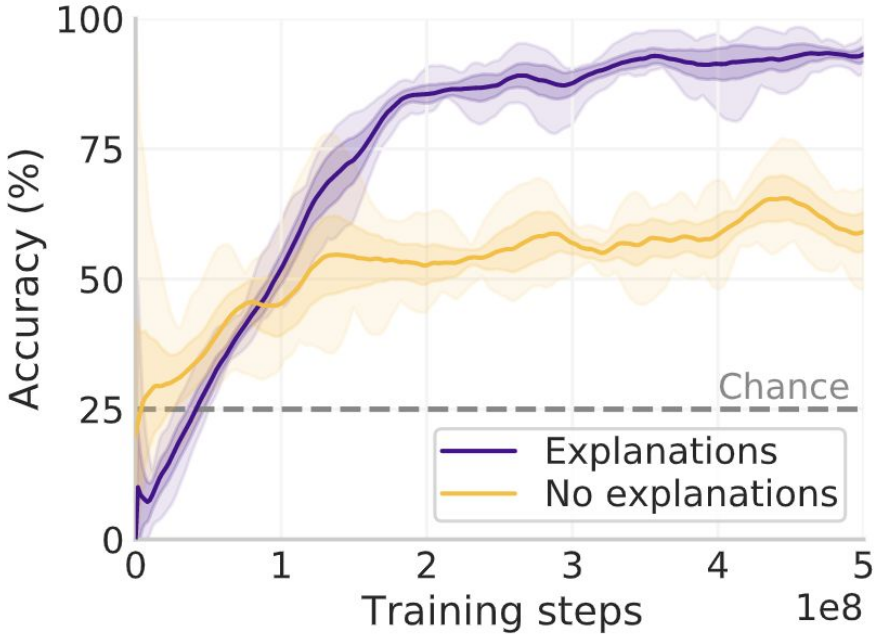
“Incorrect, other objects are large purple, striped, or pentagons .”

Predicting explanations in practice

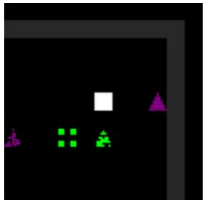


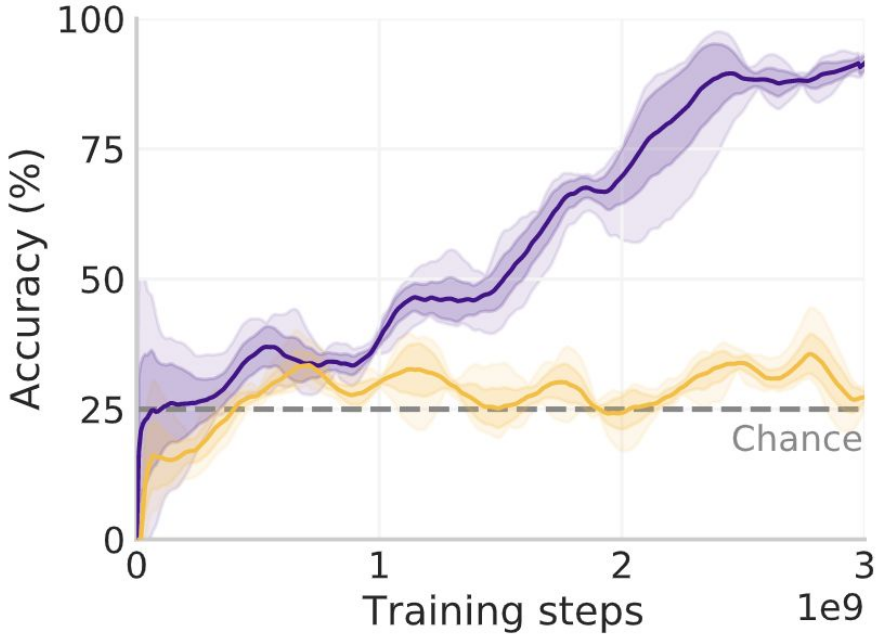


Explanations help RL agents learn odd-one-out tasks



(b) 2D results.

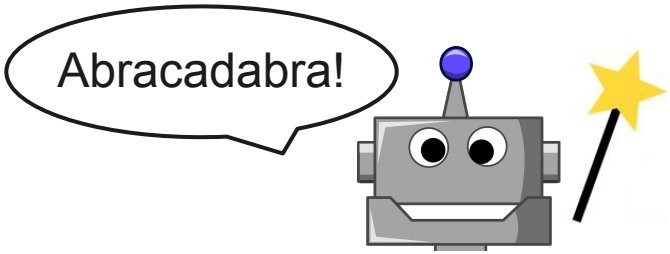




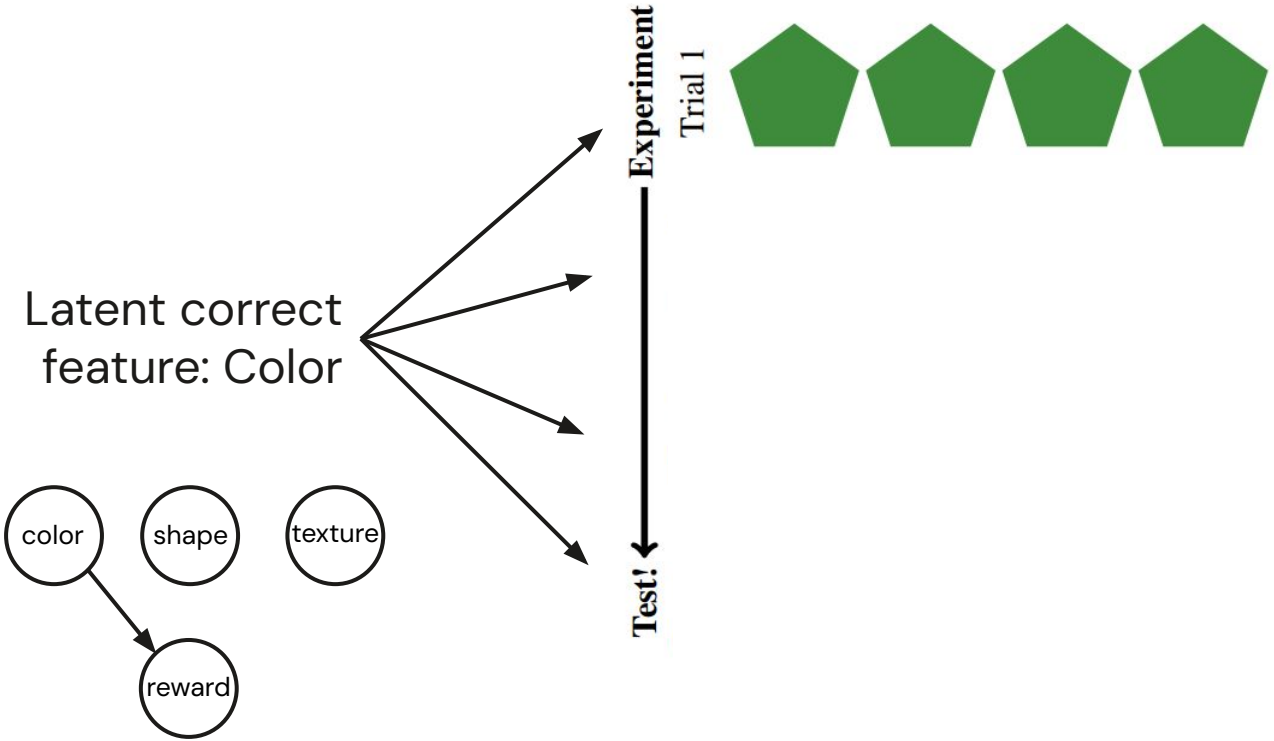
(c) 3D results.

)

Odd-one-out intervention tasks

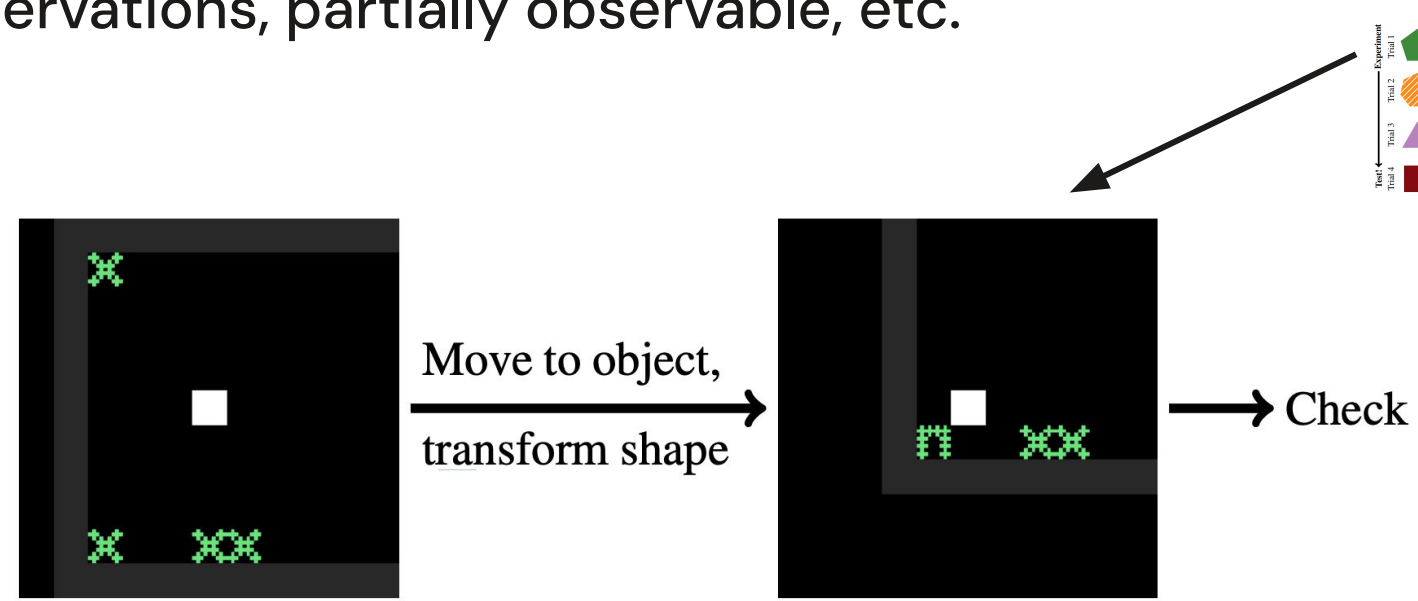


- We also considered actual experimentation + exploitation in a DAG-like task.



"Correct, the latent feature is color and it is uniquely teal."

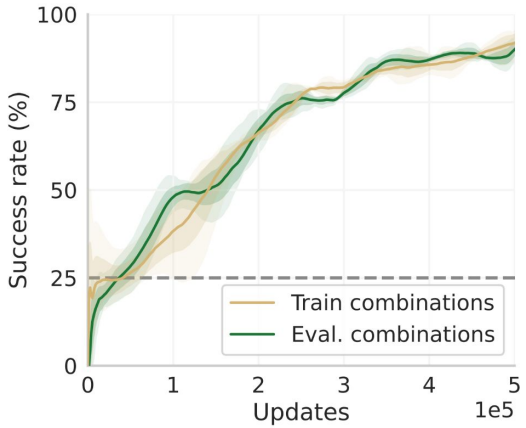
Tasks are grounded in high-dimensional observations, partially observable, etc.



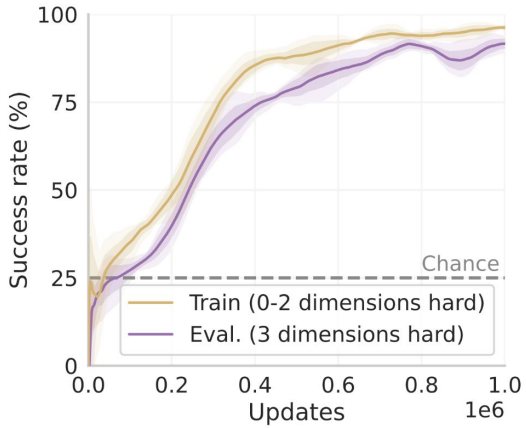
- We train the agent via BC on passive data from an expert which near-optimally acts to discover the latent structure, and then use it to achieve reward.

Agents can generalize the odd-one-out intervention tasks from passive data as well

- We did two train/test splits, either just by which feature combinations were used, or by changing experiment difficulty along different dimensions (see paper).
- In either case, agent generalizes well from passive training.

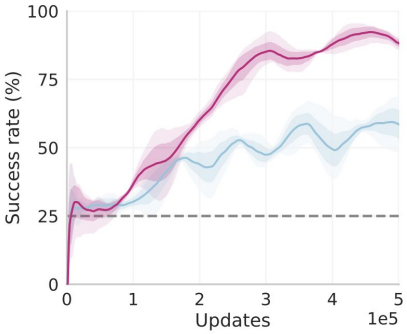


(b) Feature set generalization

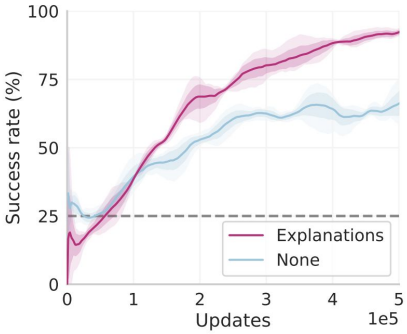


(c) Hard dimension generalization

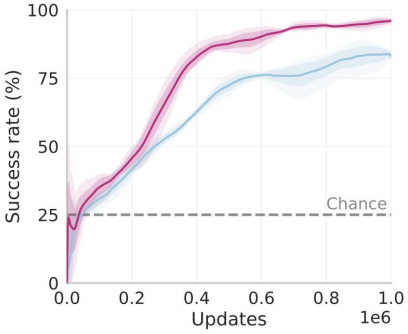
… and explanations help support that passive learning



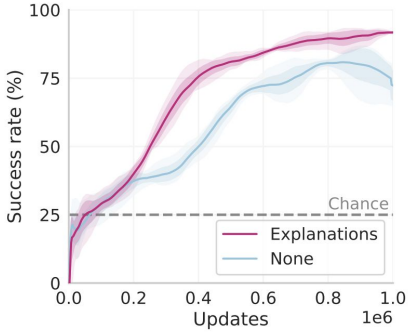
(a) Feature set generalization train.



(b) Feature set generalization eval.



(c) Hard dimension generalization train.



(d) Hard dimension generalization eval.

However, differences are not quite as dramatic as for RL agents, presumably because observing the expert policy gives more info about the task.

Agents can passively learn causal strategies in more complex environments.

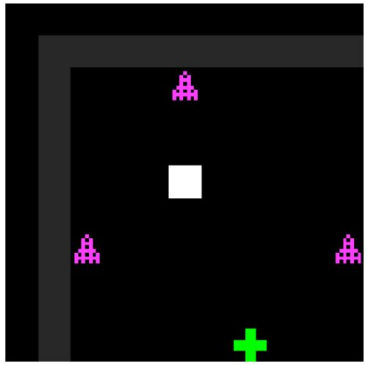
Explanations can help.

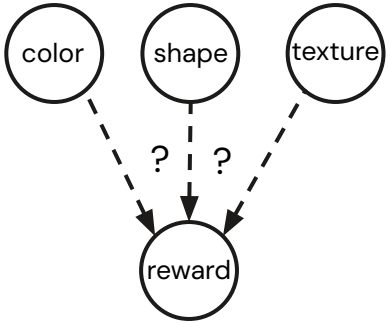
Google DeepMind

3 Explanations can shape OOD generalization

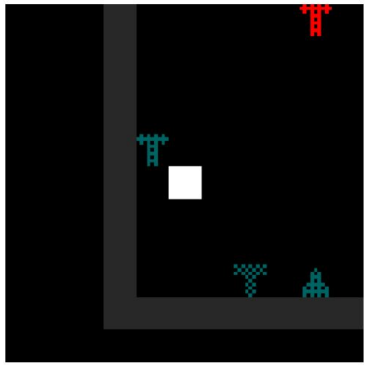
Odd-one-out confounding tasks

Train (confounded):



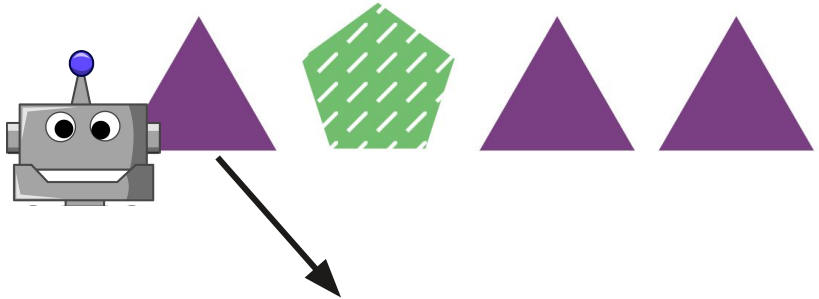


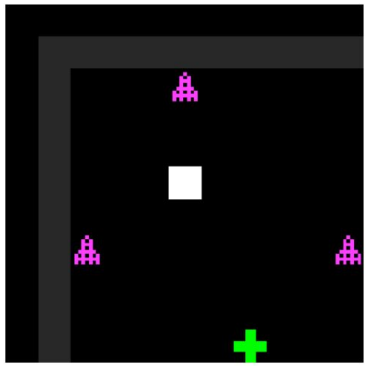
Evaluation (deconfounded):



Give explanations that focus on a specific dimension

Train (confounded):





Between-agents manipulation

Color: “Incorrect, other objects are purple.”

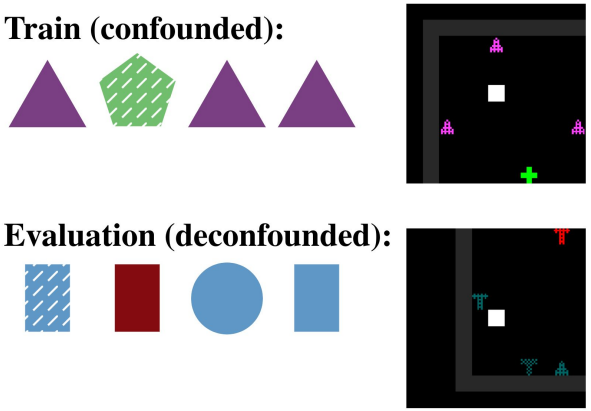
Shape: “Incorrect, other objects are triangles.”

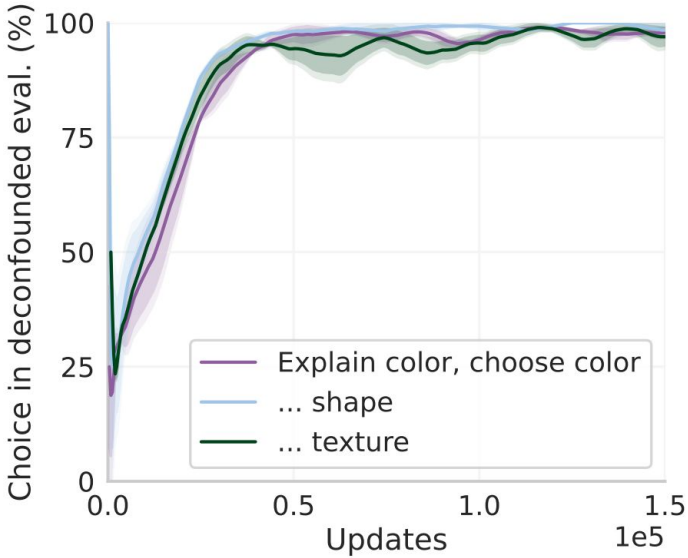
Texture: “Incorrect, other objects are solid.”

How can agent know “triangle” is the shape, not color?

Another episode:

Explanations can shape OOD generalization, even for passive learners





Agent is not forced to use these explanations for the task; features that are explained just become salient.

Predicting explanations can shape how passively-trained agents generalize

Google DeepMind

4 Language models

The internet text from which LMs learn is full of descriptions of experiments, outcomes, and explanations

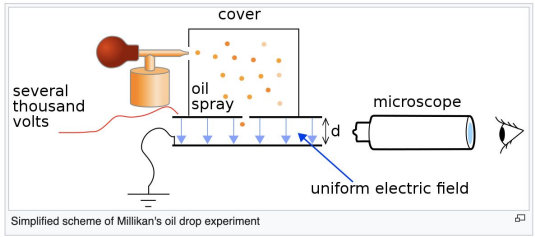
WIKIPEDIA
The Free Encyclopedia

Experimental procedure [edit]

Apparatus [edit]

> This section **needs additional citations for verification.** Please help improve this article by adding citations to reliable sources in this section. Unsourced material may be challenged and removed. *(December 2010)*
> *(Learn how and when to remove this template message)*

Millikan's and Fletcher's apparatus incorporated a parallel pair of horizontal metal plates. By applying a potential difference across the plates, a uniform electric field was created in the space between them. A ring of insulating material was used to hold the plates apart. Four holes were cut into the ring, three for illumination by a bright light, and another to allow viewing through a microscope.



Simplified scheme of Millikan's oil drop experiment

A fine mist of oil droplets was sprayed into a chamber above the plates. The oil was of a type usually used in vacuum apparatus and was chosen because it had an extremely low vapour pressure. Ordinary oils would evaporate under the heat of the light source causing the mass of the oil drop to change over the course of the experiment. Some oil drops became electrically charged through friction with the nozzle as they were sprayed. Alternatively, charging could be brought about by including an ionising radiation source (such as an X-ray tube). The droplets entered the space between the plates and, because they were charged, could be made to rise and fall by changing the voltage across the plates.

stackoverflow

I tried using a simple model:

```
model = tf.keras.Sequential([
    tf.keras.layers.experimental.preprocessing.Rescaling(1./255, input_shape=(720,
    tf.keras.layers.Conv2D(128, 5, padding='same', activation='relu'),
    tf.keras.layers.MaxPooling2D(),
    tf.keras.layers.Conv2D(128, 5, padding='same', activation='relu'),
    tf.keras.layers.MaxPooling2D(),
    tf.keras.layers.Conv2DTranspose(2, [720, 1280])
])

model.compile(optimizer='adam', loss=tf.keras.losses.SparseCategoricalCrossentropy
```

But recieve this error:

> ValueError: Input 0 of layer "sequential_24" is incompatible with the layer: expected shape=(None, 720, 1280, 3), found shape=(720, 1280, 3)

I'm pretty sure that is very easy to fix, however I couldn't manage to do so. Basically the "array size" is missing. I tried playing around with `[]` and `()` or `tf.data.Dataset.from_tensor_slices`, but the best I could achieve was of shape (2, 720, 1280, 3) where the label column was missing then...

Any idea on how to correctly set up the dataset or adjust the model?

python tensorflow

Evaluating LMs

- We wanted to test the kind of causal strategies that we explored with the agents.
- On LMs that are only trained passively on pure language modeling (unlike ChatGPT or Bard).
- So used Chinchilla (70 billion parameters).
- And turned the odd-one-out interventions task into a language-based task, complete with various kinds of explanations.
- Hopefully this is a weird enough task to not be in training text (but model will obviously be familiar with concepts like uniqueness).

There are three objects in front of me:
A) green pentagon solid
B) green pentagon solid
C) green pentagon solid
I transform object **C** into a different **texture**: striped.
Choosing object C was not rewarded.
Outcome explanation: **The rewarding dimension must not be texture.**

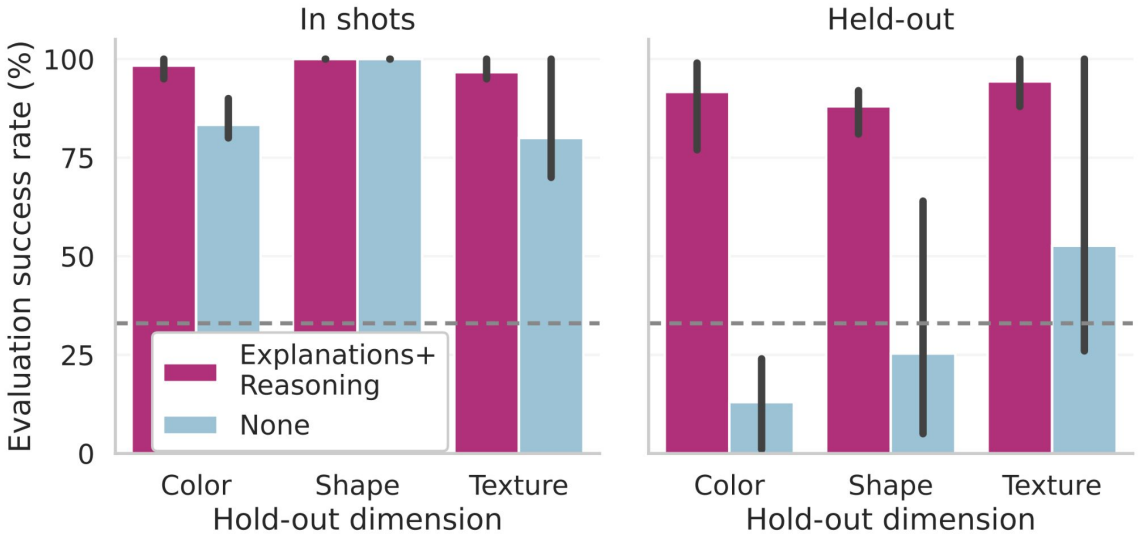
[...]

There are three objects in front of me:
A) purple ellipse solid
B) green trapezoid solid
C) green ellipse striped
Reasoning: Let's think step by step. **In this game, I am rewarded for unique shape. Object B is the only trapezoid object, because A and C are ellipses, so B has a unique shape.**
I choose object **B**
Choosing object B was rewarded!
Outcome explanation: **I was rewarded for unique shape in this game.**

(wording slightly altered for clarity and brevity)

Testing LMs on the odd-one-out interventions

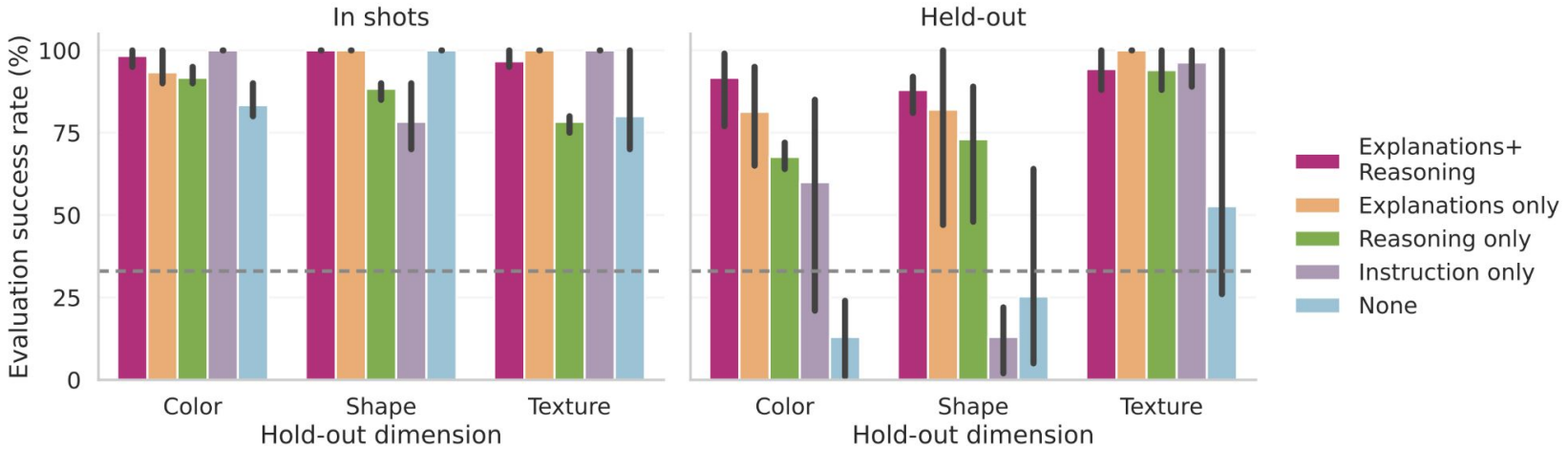
- Give Chinchilla a 4-shot prompt (4 example games) with expert choices and explanations + reasoning.
- In prompt shots are sampled such that the rewarding feature is chosen from two dimensions (e.g. color and shape), while a third (e.g. texture) is held out (not rewarding) in the prompt.
- Select the prompt automatically for performance on new tasks from the included dimensions.
- Evaluate performance on tasks from the held-out dimension.



LMs can learn the odd-one-out intervention task from examples in context.

With explanations, they can generalize.

Which kinds of explanations matter? Either outcome explanations or reasoning before choice will do



Google DeepMind

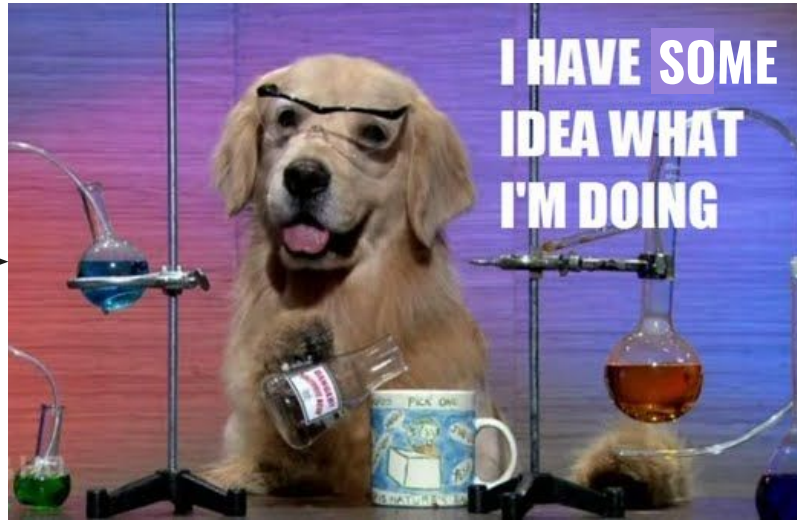
5 Wrapping up

Metaphorically, can you learn to be a scientist just by reading enough books explaining experiments?

Passive learning

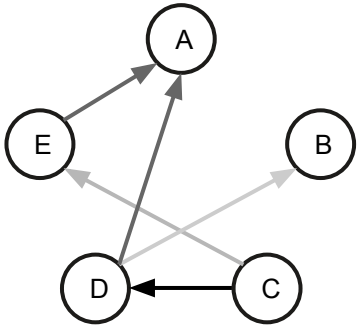
✓

Active sciencing and new discoveries!



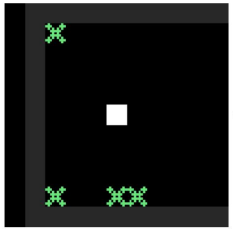
In some cases! Especially because science books often explain *why* an experiment was done, and what the results imply.

More formally



- Observational data does not generally allow learning causality
- However, it's possible to learn causal **strategies** for **actively** experimenting to determine causal structure, and then exploiting it, from passive data.
  - At least as long as that data includes examples of experts experimenting.
  - Without ever explicitly inferring or supervising any DAG.
- This works in toy causal DAG environments, and more complex ones with pixel observations and relational structure.
- Explanations can help support causal learning, and can shape generalization from confounded data.
  - Generally, language is a powerful learning signal.
- LMs can generalize causal strategies from a few-shot prompt, if that prompt includes explanations.

Train (confounded):



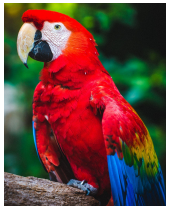
Some footnotes for prior claims on LMs and causality

"This hierarchy, and the formal restrictions it entails, explains why statistics-based machine learning systems are prevented from reasoning about actions, experiments and explanations. [...] the hierarchy denegrades [*sic*] the impressive achievements of deep learning to the level of Association [...] Unfortunately, the theoretical barriers that separate the three layers in the hierarchy tell us that the [...] objective function does not matter. As long as our system optimizes some property of the observed data, however noble or sophisticated, while making no reference to the world outside the data, we are back to level-1 of the hierarchy with all the limitations that this level entails."*

Gary Marcus ✓
@GaryMarcus

Part of this is right: GPT only knows about correlations within text usage, not about the real world. But scientists –-and children as @AlisonGopnik has emphasized--strive to derive causal models of the world, even a world inferred indirectly.

GPT never does that.*



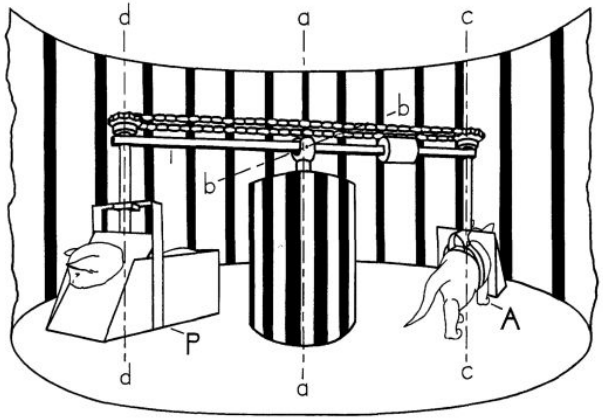
Causal Parrots: Large Language Models May Talk Causality But Are Not Causal*

MORITZ WILLIG* and MATEJ ZEČEVIĆ*, Technical University of Darmstadt, Germany
DEVENDRA SINGH DHAMI, Technical University of Darmstadt and hessian.AI, Germany
KRISTIAN KERSTING, Technical University of Darmstadt, hessian.AI and DFKI, Germany

*Actually, LMs **could** learn quite a bit about causality & experimentation from passive data!

Caveats: what this work does not imply

- Doesn't imply that passive learning is as good as active learning.
  - We know that passive learning is worse for humans, animals, and agents.
  - BC is fundamentally limited by quality of data (at least unconditional BC).
  - And even ignoring causality, active experience can be more efficient; e.g. avoid repeating things you already know.
  - Indeed, most deployed "LMs" are tuned with interactive objectives,
  - Interactive training would likely improve results, particularly in more complex environments.
  - But passive data can go a long way.
- Doesn't imply that confounding is not a problem.
  - Explanations might overcome confounding in some cases, but only if they are present and accurate, which is far from guaranteed.
  - Since even humans scientists have a hard time resolving confounding, not every explanation on the internet is right…



ELSEVIER

The American Journal of Medicine

Volume 135, Issue 10, October 2022, Pages 1213-1230.e3

Clinical Research Study

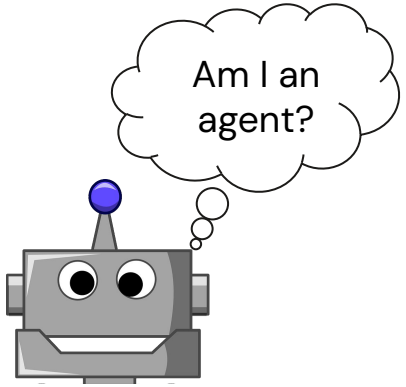
Alcohol Consumption and Cardiovascular Health

Chayakrit Krittanawong MD [a b], Ameesh Isath MBBS [c], Robert S. Rosenson MD [d e], Muzamil Khawaja MD [b], Zhen Wang PhD [d e f g], Sonya E. Fogg MLS [h], Salim S. Virani MD, PhD [a b], Lu Qi MD, PhD [i], Yin Cao MPH, ScD [j], Michelle T. Long MD, MSc [k], Christy C. Tangney PhD [l], Carl J. Lavie MD [m]

Future directions and further food for thought

- How good is the data on the internet actually for learning causal skills?
  - Lots of correct explanations, but also lots of conspiracy theories etc.
  - Our results used high-quality expert data; would be interesting to explore relaxing that assumption and incorporating noisier data.
  - As above, LMs would likely do better if fine-tuned for these capabilities on high-quality expert data.

- How does this connect to agentic, goal-directed behavior?
  - Obviously, exploring and exploiting to achieve a goal (reward) seems like agentic behavior. Agents can learn this from purely passive imitation, and LMs can be pushed into it from a prompt.
  - Is there a deep and fundamental difference between internally-driven goal directed behavior (e.g. in humans or animals) and prompted goal-driven behavior as in pretrained LMs?
  - Note that human/animal behavior can be shifted by context as well…
  - And that LMs do have “default” internally-driven behavior without a prompt (especially after tuning, but even before), though it is perhaps both less coherent and more chaotic (self-conditioning).



https://arxiv.org/abs/2305.16183
https://proceedings.mlr.press/v162/lampinen22a.html

Thanks!
Questions?