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]

Abstract

Inferring the abstract relational and causal structure of the world is a major challenge for reinforcement-learning (RL) agents. For humans, language—particularly in the form of *explanations*—plays a considerable role in overcoming this challenge. Here, we show that language can play a similar role for deep RL agents in complex environments. While agents typically struggle to acquire relational and causal knowledge, augmenting their experience by training them to predict language descriptions and explanations can overcome these limitations. We show that language can help agents learn challenging relational tasks, and examine which aspects of language contribute to its benefits. We then show that explanations can help agents to infer not only relational but also causal structure. Language can shape the way that agents to generalize out-of-distribution from ambiguous, causally-confounded training, and explanations even allow agents to learn to perform experimental interventions to identify causal relationships. Our results suggest that language description and explanation may be powerful tools for improving agent learning and generalization.

It is often argued that machine learning models—and deep learning models in particular—lack the human proficiencies for forming abstractions and inferring relational or causal structure (e.g. Fodor & Pylyshyn, 1988; Lake et al., 2017; Pearl, 2018; Marcus, 2020; Ichien et al., 2021; Holyoak & Lu, 2021; Puebla & Bowers, 2021; Geirhos et al., 2020). These limitations can make it hard to train models that generalize out-of-distribution, or that reason in human-like ways, particularly for reinforcement learning (RL) agents that receive high-bandwidth input from raw pixels and must learn to act in partially-observable environments.

[1]DeepMind, London, UK. Correspondence to: Andrew Lampinen <lampinen@deepmind.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Human learning of abstract, relational, and causal structure benefits substantially from language, and particularly *explanations*. Language helps us to identify structure in the world, and to structure our thinking (Edmiston & Lupyan, 2015; Lupyan, 2016; Dove, 2020). *Explanations*—language that provides explicit information about appropriate abstractions and causal structure (Keil et al., 2000; Lombrozo, 2006)—are particularly useful. Explanations mitigate credit assignment problems, by linking a concrete situation to reusable abstractions (Lombrozo, 2006; Lombrozo & Carey, 2006). Thus, humans can use explanations learn efficiently, from otherwise underspecified examples (Ahn et al., 1992). Explanations help us make comparisons and master relational and analogical reasoning (Gentner & Christie, 2008; Lupyan, 2008; Edwards et al., 2019). Explanations selectively highlight generalizable causal factors in a situation, improving our causal inference (Lombrozo & Carey, 2006). Even explaining to ourselves, without feedback, can improve our generalization (Chi et al., 1994; Rittle-Johnson, 2006; Williams & Lombrozo, 2010), perhaps because explanations form abstractions that are easy to recall and generalize (cf. Dasgupta & Gershman, 2021).

Indeed, there has been increasing interest in using language or explanations as a learning signal for machines (e.g., Ross et al., 2017; Mu et al., 2020; Camburu et al., 2018; Schramowski et al., 2020; see related work). That is, rather than seeking explanations post-hoc, to help humans understand a system (e.g., Chen et al., 2018; Topin & Veloso, 2019; Xie et al., 2020), these works use explanations to help a system understand a task (cf. Santoro et al., 2021). Most machine learning from explanations focuses on supervised learning[1], but explanations may be even more relevant to reinforcement learners. While supervised learners are theoretically limited (e.g., Pearl, 2018), RL agents can intervene and thus can acquire causal knowledge (e.g., Dasgupta et al., 2019; Rezende et al., 2020). Furthermore, RL agents struggle with credit-assignment, abstraction, and generalization (Ghosh et al., 2021; Kirk et al., 2021)—the exact settings where explanations help humans. These observations motivate exploring whether language explanations could help RL agents infer relational and causal structure.

[1]The few studies on explanations in RL (Guan et al., 2021; Tulli et al., 2020) do not explore relations or causal interventions.

What is a language explanation? We define a language explanation to be a string indicating a relationship between a situation, the agent's behavior, and abstract task structure. For example, after turning on an oven, an agent might receive an explanation like "turning on the oven heats it up, to prepare for baking." This explanation conveys the abstract causal links between an action and a desired goal. By contrast, many true statements would not qualify as explanations, because they ignore the agent's behavior, task-relevant structure, or both; e.g., noting that "the oven is silver" would not be an explanation unless that fact is relevant to the task. We use the term "explanation" in this work to refer to a broad class of language utterances, including descriptions if they convey task-relevant abstractions; see Discussion.

We explore the benefits of explanations using tasks situated in rich 2D and 3D environments. To study relational learning and abstraction, we use a challenging relational tasks involving uniqueness—identifying the object from a set that is the *odd-one-out* along one of multiple varying dimensions. To study causality, we first study learning in ambiguous, causally-confounded tasks, where multiple distinct features perfectly predict reward in training, but the features are dissociated in evaluation. We then study causal interventions, where agents must learn to perform experimental interventions to identify the causal structure of a particular episode.

In all settings, we find that deep RL performs poorly, but that learning and generalization improve substantially when agents learn to predict language explanations. Explanations help agents learn, by discouraging them from fixating on easy-but-inadequate "shortcut" features (Geirhos et al., 2020; Hermann & Lampinen, 2020). Explanations can help agents to disentangle confounded features, and can shape the way that agents generalize out-of-distribution on deconfounded evaluations. They also enable agents to learn to perform experiments in order to identify causal structure.

To understand these effects better, we explore how different aspects of explanations contribute to their benefits We demonstrate that the most effective way to exploit explanations is to train agents to predict them, rather than simply observe them (prediction also avoids a need for explanations in evaluation). We further show that explanation prediction is learned much more rapidly than the tasks, supporting the idea that language helps agents learn task-relevant abstractions, which in turn make learning the task easier. We furthermore show that explanations that provide feedback relevant to the specific behavior of the agent are more effective than behavior-agnostic signals, even unsupervised auxiliary objectives (input reconstruction) that provide much more information. Thus, the distinct benefits of explanations can outperform or complement more generic auxiliary learning.

Taken together, these results suggest that generating explanations could be a powerful tool for augmenting RL in challenging tasks. Furthermore, explanations posed in natural language may be simpler for humans to produce than other forms of supervision (e.g., Cabi et al., 2019; Guan et al., 2021). Thus, training agents to generate such explanations is a viable path towards both improved learning and generalization, and perhaps toward more human-like and interpretable agent behavior.

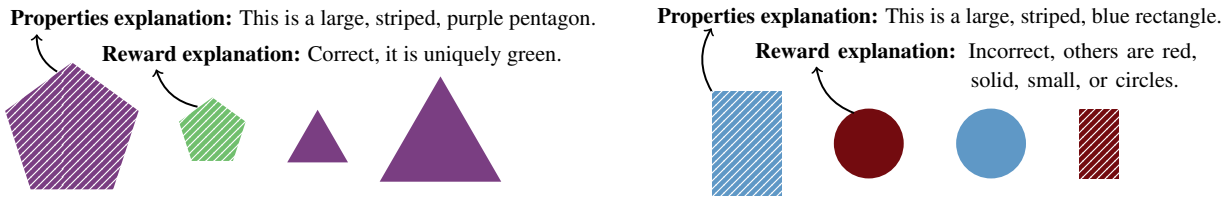
1. The odd-one-out tasks

We first outline a challenging family of fundamentally-relational tasks: finding the odd one out in a set of objects, i.e. the one that is somehow unique (Fig. 1). Odd-one-out tasks have been used extensively in cognitive science (e.g., Stephens & Navarro, 2008; Crutch et al., 2009), and proposed in perceptual settings in robotics (Sinapov & Stoytchev, 2010). These tasks are challenging, because they involve both relational reasoning (same vs. different) and abstraction (identifying uniqueness requires reasoning over all objects, and all dimensions along which objects may be related). Furthermore, these tasks permit informative explanations of relevant dimensions, properties, and relations.

Investigating these challenging and abstract—yet explainable—relational tasks is particularly interesting, because relational reasoning and abstraction are critical human abilities (Gentner, 2003; Penn et al., 2008), but the capacity of deep learning to acquire these skills is disputed (Santoro et al., 2017; 2018; Geiger et al., 2020; Ichien et al., 2021; Puebla & Bowers, 2021). However, explanations support human relational learning (Gentner & Christie, 2008; Lupyan, 2008; Edwards et al., 2019), suggesting that explanations might similarly help machines acquire these skills.

In Fig. 1 we conceptually illustrate some odd-one-out tasks. In Fig. 1a one object is uniquely green, while the rest are purple. We thus denote color as the *relevant* dimension in this episode. Along the other, irrelevant dimensions—shape, texture, and size—attributes appear in pairs, e.g. there are two pentagons and two triangles. These pairs force the agent to consider *all* the objects. If the agent considered only the first three objects it would be unable to tell whether the first object was the odd one out (uniquely large), the second (uniquely green), or the third (uniquely a triangle or uniquely solid textured). Thus, the agent must consider all objects to identify the correct dimension and the unique feature. This makes the relational reasoning particularly challenging, since the agent must consider many possible relationships. The agent is rewarded for selecting the odd-one-out, by picking it up or touching it.

We emphasize that in principle these tasks can be learned from reward alone—language is not necessary for performing them, and we evaluate without language. Nevertheless, we find that in practice language explanations are critical



(a) Example 1. (b) Example 2.

Figure 1: Conceptual illustrations of two possible odd-one-out tasks, and corresponding possible explanations. This figure depicts odd-one-out tasks with feature dimensions of color, texture, shape, and size, and the two types of explanations we consider. Property explanations identify relevant object features, while reward explanations specify which feature(s) make the choice correct or incorrect. (a) The second object is the odd one out, because it is a unique color. (b) The first object is the odd one out, because it is uniquely large. Explanations of incorrect choices identify all features.

for learning these tasks in our settings. We consider two types: reward explanations and property explanations (see Fig. 1). Reward explanations are produced after the agent chooses, and identify the feature(s) that make the choice correct or incorrect. Property explanations are produced before the agent chooses, and explain the identity of the object the agent is facing by specifying its task-relevant properties. Both types satisfy our criterion for explanations: they link the situation and the agent's behavior to the task structure.

Environments: Odd-one-out tasks can be instantiated in various settings, from games to language or images, and can incorporate various latent structures (e.g. meta-learning). We instantiate these tasks in 2D and 3D RL environments (Fig. 3a). In 2D, the agent has simple directional movement actions, while in 3D it can move, look around, and grasp nearby objects at which it is looking. In both environments we place an agent in a room containing four objects, which vary along feature dimensions of color, texture, position, and either shape (2D) or size (3D). In each episode, one object will be unique along one dimension. The 3D environment compounds the difficulty of the odd-one-out tasks, because the agent's limited view often forces it to compare objects in memory. See Appx. C.2 for full details.

2. Method: generating explanations

We focus on language explanations provided by the environment during training. We synthetically generate the explanations online, conditional on agent behavior. However, explanations could be produced by humans, e.g. as annotations of past trajectories (cf. Ross et al., 2017). We train the agent to predict explanations as an auxiliary signal to shape its representations (Fig. 2), as opposed to providing explanations as direct inputs (which is less effective; Appx. A.3); our approach thus does not require explanations at test time. Note that we do not directly supervise behavior through explanations, nor tell the agent how to use them. The agent simply predicts explanations as an auxiliary output.

We train agents using the IMPALA (Espeholt et al., 2018)

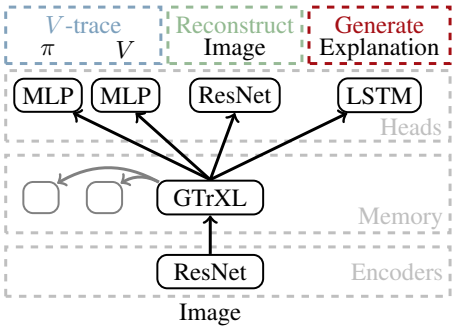


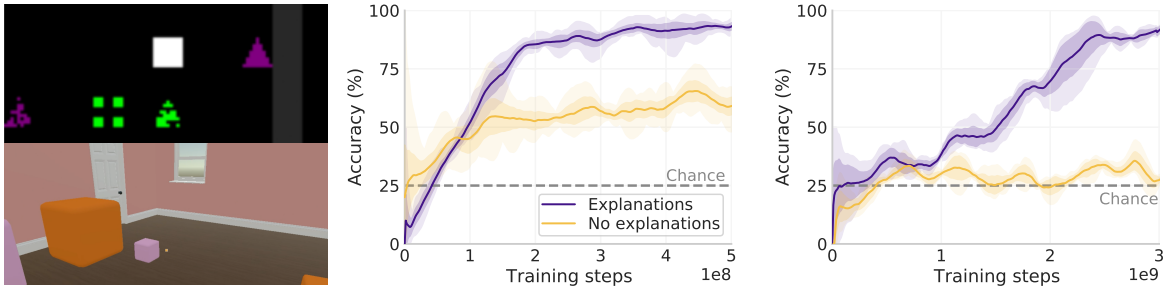
Figure 2: RL agent with auxiliary explanation prediction.

framework. Our agent (Fig. 2) consists of a visual encoder, a memory, and output heads. The encoder is a CNN or ResNet (task-dependent). The agent memory is a 4-layer Gated TransformerXL Memory (Parisotto et al., 2020), which receives the visual encoder output and previous reward as inputs. The output of the memory is input to the heads. The policy and value heads are MLPs, trained with V-trace. Another head reconstructs the input images to learn better representations (though this is not necessary; Appx. A.8). Finally, the explanation head is a single-layer LSTM, which generates language explanations. We train the agent to predict these explanations using a summed softmax cross-entropy loss. See Appx. C.1 for further agent details.

3. Experiments

3.1. Odd-one-out tasks in 2D and 3D RL environments

We first evaluate the benefit of explanations for learning odd-one-out tasks, by comparing agents trained with property and reward explanations to agents trained without. In both 2D and 3D environments, agents trained with explanations learn to solve the tasks over 90% of the time (Figs. 3b-c). Agents trained without explanations perform worse; in the easier 2D environment they exhibit partial learning (see 3.4), while in the challenging 3D environment they perform near chance. In 2D all agents were trained with an unsupervised reconstruction loss. However, agents trained



(a) 2D & 3D environments. (b) 2D results. (c) 3D results.

Figure 3: Explanations help agents learn the perceptual odd-one-out tasks in both RL environments. (a) Our environments in 2D (top) and 3D (bottom). In 2D, the agent is the white square, while in 3D it has a first-person view. The objects appear in varying positions, colors, textures, and shapes (2D) or sizes (3D). (b) 2D results. Agents trained with explanations achieve high performance; agents trained without explanations do not. (c) 3D results. Only agents trained with explanations exhibit learning substantially above chance. (Training steps denotes actor/environment steps, number of parameter updates is $\sim 10^4\times$ smaller. 5 seeds per condition in 2D, 3 per in 3D, lines=means, dark region=$\pm$SD, light region=range.)

without reconstruction but with explanations perform well (Appx. A.8), while agents trained with reconstruction but without explanations do not. By highlighting abstract task structure, explanations outperform task-agnostic auxiliary objectives, even ones that provide strictly more supervision.

3.2. Explanations can deconfound

For humans, explanations help identify *which specific aspects* of a situation are generalizable (Lombrozo & Carey, 2006). Could explanations also help RL agents to disentangle causally-confounded (perfectly correlated) features, and shape how agents generalize to out-of-distribution tests? We explore this with a different training and testing setup (Fig. 4a). In training, one object is the odd-one-out along *three* feature dimensions (color, shape, and texture). Thus, any or all of these features could be used to solve the task—the dimensions are perfectly confounded. In test, however, the features are deconfounded: there is a different odd-one-out along each dimension. We explore the effect of explanations that consistently refer to a single feature dimension (without mentioning others) on the agent's behavior in deconfounded evaluation. We train agents in four conditions: no explanations, color-only explanations, shape-only explanations, or texture-only explanations. Single-dimension explanations can potentially draw the agent's attention to a particular dimension, and thereby disentangle these features,[2] even though the explanations do not alter the relationship between these dimensions and the reward signal.

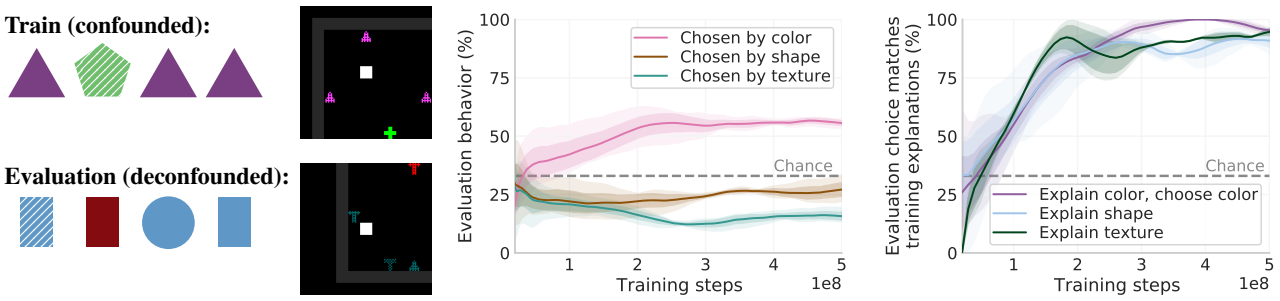
Agents trained without explanations were biased towards using color (the simplest feature) in the deconfounded evaluation (Fig. 4b). However, the agents trained with explanations generalized in accordance with the dimension that they were trained to explain $> 85\%$ of the time (Fig. 4c), even though there were no direct cues linking the reward to that dimension over the others. In this setting, shaping an agent's internal representations through explanations draws its attention to the desired dimension, and allows $> 85\%$ out-of-distribution generalization along that dimension.

3.3. Explanations allow agents to learn to experiment

Explanations help humans to understand causal structure (Lombrozo, 2006; Lombrozo & Carey, 2006). The ability of deep learning to infer causality is sometimes questioned (e.g. Pearl, 2019), but while theoretical limitations hold for passive learners, RL agents can intervene and can therefore identify causal structure. Indeed, agents can meta-learn causal reasoning in simple settings (Dasgupta et al., 2019) where causal variables are directly observable and actions allow strong interventions. We investigate whether explanations could help agents learn to identify causal structure in more challenging relational tasks in richer environments.

We consider a meta-learning setting where agents complete episodes composed of four odd-one-out trials. In each episode, there is only one causally-important dimension in all four trials—reward is determined by uniqueness along only one of the feature dimensions (e.g. color). This "correct" dimension changes across episodes, and is not directly observable. Thus agents must learn to *perform experiments* on the first three trials to identify the causally-relevant dimension, in order to select the correct object on a fourth test trial (Fig. 5a). The agent receives 1 reward for completing an early trial correctly, but 10 reward for completing the final trial correctly. Thus, the agent is incentivized to experiment and discover the correct dimension in the early trials,

[2]Feature uniqueness is always confounded, but feature values recombine across episodes, allowing disentangling.



(a) Training & evaluation setup. (b) Bias without explanations. (c) Choices with explanations.

Figure 4: Explanations can deconfound perfectly correlated features. (a) Schematic depictions and environment screenshots from train and test. The agent is trained in confounded settings, where the target object is unique in color, shape, and texture. The agent is tested in deconfounded settings, where one object is unique along each dimension (and an additional distractor object has no unique attributes). (b) When trained without explanations, the agent is biased towards using color (the simplest feature) in evaluation. (c) However, if the agent is trained with explanations that target any particular feature, the agent prefers that feature in the deconfounded evaluation. (3 seeds per condition, chance is random choice among valid objects.)

in order to gain a large reward in the final trial.

To enable experiments, in the first three trials of each episode, we give the agent a magic wand that can perform one causal intervention per trial: changing an object's color, shape, or texture. That is, we endow the agent with three additional actions which transform one of those three properties of an adjacent object. The agent is forced to use the wand to create an odd-one-out, because each trial's initial configuration lacks any objects with unique features—along each dimension the features are either all the same, or appear in pairs. When the features are all the same, the experiments are relatively easy (Fig. 5b): the agent must simply transform an object and then select the same object. When the features are paired, however, the experiments are harder (Fig. 5c): the agent must transform one object, which will change to match other objects *and then it must select another object* that was *formerly* paired with this one, but is now unique. The final trial is always a deconfounded test, where a different object is unique along each dimension, and the magic wand is disabled. On all trials, we reward the agent only if it selects an object which is unique along the "correct" dimension. Thus, the agent cannot reliably choose correctly unless it has already experimented with the magic wand to infer the correct dimension.

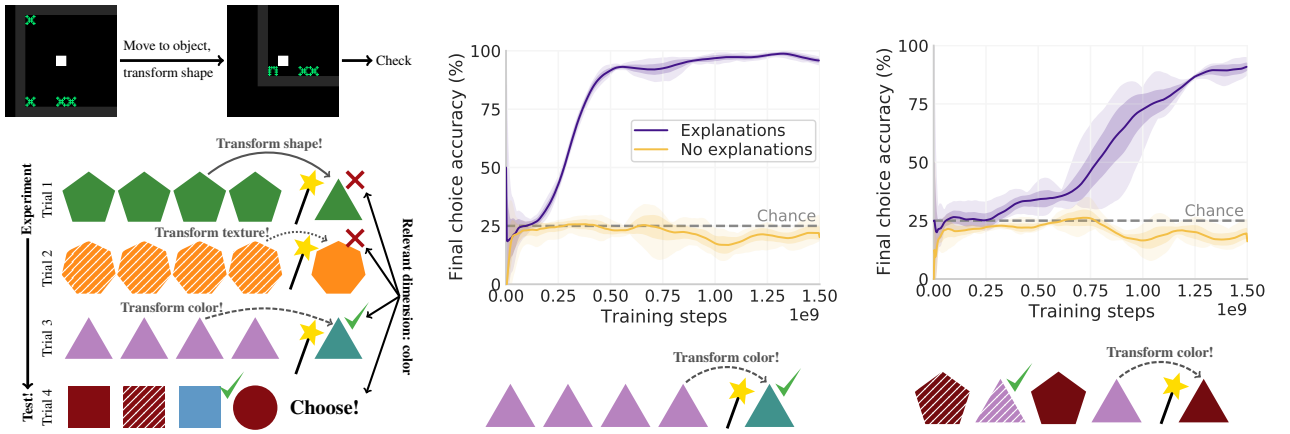
We again compare agents that receive property and reward explanations to agents that do not, but in this case the explanations are augmented to identify the correct dimension (e.g., "incorrect, the dimension is shape, and other objects are squares"). Again, while in principle these tasks could be learned from rewards alone, we find that agents trained without explanations cannot learn these tasks. However, agents trained with explanations achieve high success at both easy (Fig. 5b) and hard levels (Fig. 5c). Explanations can help agents learn to perform causal intervention experiments.

3.4. Exploring the benefits of explanation in more detail

In order to better understand the benefits of explanations, we explored our results further in a variety of analyses, ablations, and control experiments. We highlight three intriguing results here, and briefly outline the rest.

Explanations help agents overcome biases toward easy features (Appx. A.1): In 2D, agents without explanations fixate on positions and colors, and learn to solve the task only when those dimensions happen to be relevant. Shape and texture are generally not learned at all. This explains the moderate performance without explanations. With explanations, by contrast, agents learn to solve the tasks with any feature. Similarly, in the confounded features setting color is preferred without explanations, but again with explanations agents can learn to use other features. Hermann & Lampinen (2020) show similar feature-difficulty rankings for CNNs, and that CNNs lazily prefer easier features. Similarly, Geirhos et al. (2020) discuss "shortcut features" that networks prefer, despite the fact that those features do not correctly solve the task. Thus, explanations may help an agent to overcome biases towards easy-but-inaccurate solutions, to escape minima or plateaus, and to master the task.

Both explanation types provide complementary benefits; their relative value depends on the environment (Fig. 6): In the above experiments we provided agents with both property and reward explanations. Here, we compare to agents trained to generate only a single type of explanations. We found that having both types of explanations is generally better than (or at least as good as) a single type, but the relative benefits of different types depend on the setting. In the 2D environment (Fig. 6a) either type of explanations alone results in learning, but both types together result in substantially faster learning. In the 3D setting (Fig. 6b), we find that



(a) A single intervention trial (top), and multi-trial episode structure (bottom).

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

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

Figure 5: Explanations allow agents to meta-learn to perform experiments. (a) Each episode consists of four trials: three where the agent gets to experiment with a magic wand in order to discover which feature dimension is relevant, followed by a final deconfounded trial where it must choose the unique object along that dimension. In this case the relevant dimension is color. In the first trials the agent transforms the shape and texture of the objects, but is not rewarded for picking them up (red X). In the third trial, it transforms the color and is rewarded for picking the object up (green check). The agent can then infer that it should choose the different-colored object in the final trial. (b) In some episodes, the experiments are easy, because all the object attributes are the same, and the agent only needs to transform an object and select that object. Agents trained with explanations learn these tasks, while agents trained without explanations do not. (c) In other episodes, the experiments are harder, because the object attributes are all paired—the agent must transform one object, and then pick up *another* which has been made unique. With explanations, agents learn these difficult levels as well. (4 seeds per condition.)

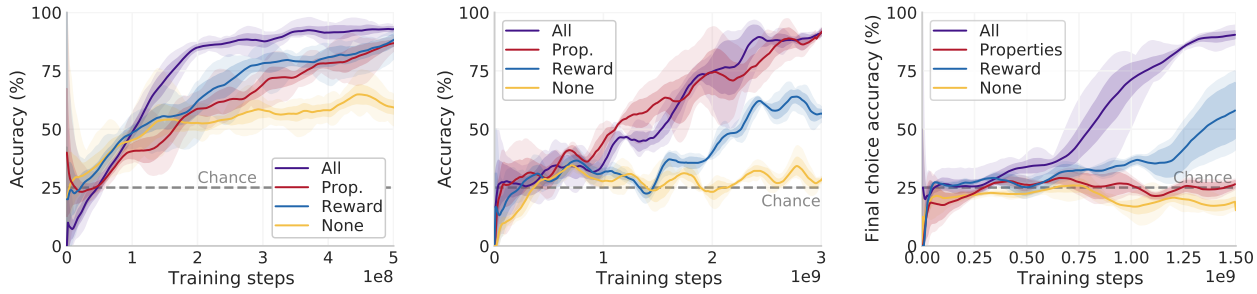
property explanations are relatively more beneficial; perhaps because predicting explanations on encountering the objects helps the agent overcome the memory challenges in the 3D environment by helping it to encode the relevant features in an easily decodable way. By contrast, for the meta-learning tasks (Fig. 6c), we find that the reward explanations are necessary for any learning. This is likely because the relevance of a transformation experiment to the final reward is much more directly conveyed by the reward explanations than the property ones. However, both types of explanations together are required for complete learning within the training budget we considered. In summary, the relative benefits of the explanation types depend on the demands of the environment, but generally having both types is best.

Behaviorally- and contextually-relevant explanations are best (Appx. A.2): Human explanations are *pragmatic* communication: they depend on context, knowledge, and behavior (Van Fraassen, 1988). We therefore compared to ablation explanations that referred to objects in the room, but independent of behavior (on 10% of steps we randomly chose an explanation that could occur in the current room, regardless of agent actions), and irrelevant explanations (randomly sampled from those possible in *any* room). We found that behavior-relevant explanations were much more beneficial than behavior-irrelevant ones, and completely irrelevant explanations had no benefit. In particular, in the challenging experimentation tasks, behaviorally relevant explanations were *necessary* for learning. Explanations should engage with the agent rather than passively conveying information.

Other explorations: We briefly describe our other explorations here; see Appx. A for full results. We find that language prediction is learned much faster than the RL tasks (Appx. A.6), supporting our suggestion that language makes learning abstractions easier, which can in turn support RL. We found that explanations as input are not helpful (Appx. A.3), and can even interfere with the benefits of explanations as targets. Prediction is a more powerful signal for learning than receiving an input. We also found that a curriculum of tasks that teaches the agent about object properties—by cueing the agent with a property as input (e.g. "blue") and then rewarding the agent if it chooses the matching object—is not as effective as explanations (Appx. A.7). Finally, explanations are more beneficial in complex tasks (Appx. A.5). Thus, explanations may be especially useful as RL is implied to increasingly complex settings.

4. Related work

Language plays a critical role in human learning. Language can identify consistent abstractions or structures in the world, and can shape reasoning processes (Edmiston & Lupyan, 2015; Lupyan, 2016; Dove, 2020). In particular, ex-



(a) Basic 2D. (b) Basic 3D. (c) Meta-learning (hard levels).

Figure 6: Different explanation types offer complementary, separable benefits. We compare agents trained with all explanations or none (as above) to those trained with only property explanations (red), or only reward explanations (blue). (a) In the basic 2D tasks, either kind of explanations is sufficient for learning, but having both types together is substantially faster. (b) In the 3D tasks, property explanations result in comparable learning, while reward explanations are not as effective, but still better than none. This is likely due to the memory challenge of these tasks, since it is harder to see all objects at once—property explanations can help the agent discover what to encode to make its choice, while reward explanations cannot. (c) In the learning to experiment setting, by contrast, only reward explanations result in any learning on the hard levels, but both types together is even better. (5 seeds each for All/None conditions, 2 seeds each for properties/reward.)

planations can enable efficient, generalizable learning, even from a single example (Ahn et al., 1992). Explanations can highlight both causal factors, and relationships between a present situation and broader principles (Lombrozo, 2006). Explanations therefore depend strongly on prior knowledge, and the relationship between explainer, the recipient, and the situation to be explained (Van Fraassen, 1988; Cassens et al., 2021). As Wood et al. (1976) say: "one must recognize the relation between means and ends in order to benefit from 'knowledge of results.'" Explanations link a specific situation to more general principles that can be used in the future.

Relations: Relational and analogical reasoning are considered crucial to human intelligence (Gentner, 2003), and possibly absent in other animals (Penn et al., 2008). The relations *same* and *different* are central to many accounts, but their origins are disputed (Penn et al., 2008; Katz & Wright, 2021, e.g.). But language and culture play a critical role in learning these concepts and skills (Gentner & Christie, 2008; Lupyan, 2008)—"relational concepts are not simply given in the natural world: they are culturally and linguistically shaped" (Gentner, 2003). Thus, explanations may be particularly key, and their absence may help explain neural networks' deficits in relational reasoning (Geiger et al., 2020; Puebla & Bowers, 2021; Ichien et al., 2021), at least without relational inductive biases (Santoro et al., 2017; Shanahan et al., 2020).

Causality: Humans focus on causal structure, even as children (Gopnik et al., 1999; Gopnik & Sobel, 2000), and our causal understanding is closely linked to explanations (Lombrozo & Vasilyeva, 2017). Human explanations are not just causal, but emphasize important causal factors that are useful for future prediction and intervention (Lombrozo & Carey, 2006). Furthermore, Lombrozo & Carey (2006) emphasize that children accept various explanations, while adults selectively endorse causally generalizable ones, suggesting that this focus may be at least partly learned.

Self-explanation: Asking humans to produce explanations for themselves, without providing feedback, can improve generalization (e.g. Chi et al., 1994; Rittle-Johnson, 2006; Williams & Lombrozo, 2010). Furthermore, Nam & McClelland (2021) find that the ability to produce explanations is strongly related to the ability to learn a generalizable problem-solving strategy involving relational reasoning; and furthermore that education—especially in mathematics—is related to developing these abilities. The skills of explaining and generalizing may be learned together.

4.1. Related work in AI

We are certainly not the first to observe that the cognitive literature suggests that explanations might help in AI. Here we review a variety of prior work on explanation in AI. We also relate to the broader set of approaches for auxiliary supervision that help agents (or models) to learn more effective representations for a task. Explanations are a particularly targeted form of auxiliary supervision that focuses on the causally-relevant, generalizable elements of a situation.

Language as representation, or to shape representations? Andreas et al. (2018) used language as a latent bottleneck representation in meta-learning, and found benefits. However, Mu et al. (2020) showed that it was better to *not* bottleneck through language, but merely use descriptions to shape latent representations in supervised classi-

fication tasks. We similarly use language as an auxiliary signal to shape latent representations toward task-relevant abstractions. However, we focus on RL, where discovering the right abstractions is more challenging and therefore language can be even more beneficial. RL also allows us to extend to settings like causal intervention, which is not possible in a classification paradigm.

Natural Language Processing: Explanations fit naturally into NLP tasks, and Hase & Bansal (2021) highlight the many ways that explanations could enter in NLP tasks, e.g. as targets, inputs, or as priors. They find no improvement from using explanations as targets, but show some positive effects of explanation retrieval during both training and test, including improved performance on relational tasks and better disentangling of causal factors.

Feature explanations as learning tools: Some prior work has refined models using input attention or gradients as targets for explanatory feedback (e.g. from humans). Ross et al. (2017) show that penalizing gradients to irrelevant features can improve generalization on a variety of image and language tasks. Lertvittayakumjorn & Toni (2021) survey works on tuning NLP models using explanatory feedback on features, word-level attention, etc. Schramowski et al. (2020) highlight an intriguing interactive-learning-from-feedback setting where an expert in the loop gives feedback which can be used for similar counter-example- or gradient-based training. Stammer et al. (2021) extend this approach in neurosymbolic models to intervene on symbolically-conveyed semantics rather than purely visual features. In RL, however, applications of feature explanations have been more limited, although Guan et al. (2021) used human annotations of relevant visual features (and binary feedback) to generate augmentations that varied the task-irrelevant features, and showed benefits over other feature-based explanation techniques or augmentations in video game playing.

Language in RL: Language is used broadly in RL, whether as instructions (e.g. Hermann et al., 2017; Kaplan et al., 2017), to target exploration (Goyal et al., 2019; Watkins et al., 2021; Mu et al., 2022; Tam et al., 2022), or as an abstraction to structure hierarchical policies (Jiang et al., 2019). Luketina et al. (2019) review many recent uses of language in RL, and argue for further research. However, they do not even mention explanations. Tulli et al. (2020) consider natural language explanations of actions in RL. However, they only evaluate a simple, symbolic MDP, and observe no benefits, perhaps because their explanations do not relate to the abstract task structure.

Auxiliary tasks: Predicting explanations is part of the general paradigm of shaping agent representations with auxiliary signals (e.g. Jaderberg et al., 2016). However, explanations are fundamentally different from unsupervised losses—unsupervised objectives are task-independent by definition, while explanations selectively emphasize the causally relevant features of a situation, and the relationship to general task principles (Lombrozo & Carey, 2006; Lombrozo, 2006). Some *supervised* auxiliary objectives are more similar to explanations; the boundaries of explanation are blurry. In the Alchemy environment (Wang et al., 2021), which involves learning latent causal structure, predicting task-relevant features improves performance. Similarly, Santoro et al. (2018) show that predicting a "meta-target"—an abstract label encoding some task structure—improves learning of a relational reasoning task. More broadly, supervising task inference can improve meta-learning (Rakelly et al., 2019; Humplik et al., 2019). Since these predictions directly relate to task structure, they are closer to explanations than unsupervised task-agnostic predictions. However, they do not necessarily actively link the details of the present situation to the principles of the task, as human explanations do.

5. Discussion

We explored relational and causal tasks that are challenging for RL agents to learn from reward alone. In all cases, learning to generate language descriptions and explanations significantly improved performance. Even though our agents lacked prior knowledge of language, they were able to rapidly learn to predict language explanations, and this prediction helped them to discover the reasoning processes necessary for the task. Explanations help agents learn the challenging, essential abilities of relational and causal reasoning.

We particularly emphasize the causal benefits of explanations, because causal understanding is essential to effective generalization. Indeed, explanations shaped how our agents to generalize out-of-distribution (cf. Ross et al., 2017) from ambiguous, causally-confounded data; furthermore, explanations enabled agents to learn to perform their own experimental interventions to identify causal structure. Without access to a pre-specified, discrete causal diagram (e.g. Pearl, 2019), our agents were able to ground explanations in pixel-level inputs to learn and generalize causal structure. Humans use explanations to highlight the causally-relevant factors of a task (Lombrozo & Carey, 2006), and our results show that explanations can play that same role for agents.

This focus on task-specific structure allows explanations to outperform task-agnostic auxiliary objectives. Indeed, we found that explanations helped agents to move beyond a fixation on easy shortcuts that do not fully solve the task, but that models nevertheless prefer (cf. Hermann & Lampinen, 2020; Geirhos et al., 2020). Explanations offer a promising route to training RL agents that learn and generalize better.

Criteria for explanations: Explanations should satisfy certain criteria for maximal benefits. Explanations must relate between the context, the agent's behavior, and the abstract

task structure—explanations that ignore behavior are less useful, and those that ignore context are useless. Receiving explanations as input was not useful in our experiments, likely because it is easier for the agent to ignore inputs than auxiliary targets. Furthermore, explanations outperform unsupervised auxiliary reconstruction. Thus, simply training agents with more information (as with unsupervised objectives) is often not sufficient; explanations must provide relevant and specific learning targets to be most beneficial.

Nevertheless, we acknowledge that the boundaries of explanation are vague (cf. Van Fraassen, 1988; Woodward, 2005; Brenner et al., 2021). For example, descriptions cannot name every property, so they tend to pragmatically focus on causally-relevant ones, and thus highlight similar features to explanations. This fact is why we refer to task-relevant property descriptions as explanations. Furthermore, we use "explanation" to refer to cues to relationships between specific situations, behaviors and abstract principles, which may overlap with other forms of auxiliary supervision. While we focused on language explanations, non-language predictions that highlight abstract task features could likely serve the same purpose. Explanations can also vary in abstraction (cf. Fyfe et al., 2014; Watkins et al., 2021). The boundaries of explanation should be explored further in future work.

Limitations and future directions: We performed our experiments on variations of odd-one-out tasks in RL, which may seem to limit the breadth of our conclusions. However, our different experiments cover many central challenges, including relational and causal structure, confounding, and meta-learning, as well as 2D and 3D environments. Thus we expect that our results will generalize to other challenging settings in future work; even beyond RL.

In broader settings, the ground truth for explanations may not be automatically accessible. In some cases, knowledge about abstractions such as property descriptions may be accessible through large pretrained models—indeed, in subsequent work we have shown that the abstractions from large language-supervised image captioning systems can improve RL exploration (Tam et al., 2022)—which may provide a scalable route to language supervision. However, these models may not have access to causal or task-specific structure, or may not work well in domains like Atari which are outside their training distribution. In such cases it may be necessary to rely on alternative approaches, particularly human annotation. Nevertheless, as noted above, explanation is an especially natural and rich form of feedback for humans to provide, so collecting such data may be worthwhile.

Although we drew inspiration from human uses of explanation, our agents do not learn from explanations in the same way as humans. Humans can use our prior knowledge of language to learn from a single explanation in context. By contrast, our agents needed to learn about language *simultaneously* with learning about the tasks, through many repetitions of similar explanations. Future work should explore whether agents that are trained with language and explanations across a diverse array of tasks can meta-learn how to learn from explanations in a more human-like way.

We also do not want to imply that explanations are *necessary* for learning. Most of our tasks could potentially be learned with sufficient data alone, especially if combined with more complicated techniques, for example data augmentation (Raileanu et al., 2020; Guan et al., 2021), or auxiliary generative model learning (Gregor et al., 2019). Furthermore, many promising domains for deep learning—such as protein structure prediction (Jumper et al., 2021)—are precisely those areas that humans do not understand well, and so are challenging domains for humans to explain. Indeed, some domains might be irreducibly complex; in these domains forcing a system to strictly follow simple explanations could be detrimental. Our approach does not force the agent to use explanations directly, and therefore might be less harmful in such cases than stronger constraints like requiring symbolic representations (e.g., Garcez & Lamb, 2020).

In other domains there may exist simple explanations that humans have not yet discovered. This observation motivates a future research direction: learning to explain over diverse task distributions, leveraging human explanations in domains we do understand. A curriculum focused on producing explanations could potentially yield substantial benefits. Humans generalize better after explaining, even without feedback (Chi et al., 1994; Rittle-Johnson, 2006), and this ability may be learned through education (cf. Nam & McClelland, 2021). An agent that similarly learns to produce explanations might similarly learn to generalize better *even in some domains for which we lack ground truth explanations*, and its explanations might help humans interpret its behavior, and the domains in which it performs.

Conclusions: We considered a challenging set of relational and causal tasks, and showed that learning to predict language descriptions and explanations helps RL agents to learn and generalize these tasks across various settings and paradigms. Explanations can help agents move beyond biases favoring easy features, determine how agents generalize out-of-distribution from ambiguous experiences, and allow agents to meta-learn to perform experiments to identify causal structure. Because these abilities are challenging for current agents, generating explanations as an auxiliary learning signal—rather than purely for post-hoc interpretation—may be a fruitful direction for further research.

Acknowledgments

We thank Antonia Creswell and MH Tessler for comments and suggestions, as well as several anonymous reviewers.

References

Ahn, W.-k., Brewer, W. F., and Mooney, R. J. Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2): 391, 1992.

Andreas, J., Klein, D., and Levine, S. Learning with latent language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2166–2179, 2018.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Brenner, A., Maurin, A.-S., Skiles, A., Stenwall, R., and Thompson, N. Metaphysical Explanation. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

Cabi, S., Colmenarejo, S. G., Novikov, A., Konyushkova, K., Reed, S., Jeong, R., Zolna, K., Aytar, Y., Budden, D., Vecerik, M., et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *arXiv preprint arXiv:1909.12200*, 2019.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31:9539–9549, 2018.

Cassens, J., Habenicht, L., Blohm, J., Wegener, R., Korman, J., Khemlani, S., Gronchi, G., Byrne, R. M., Warren, G., Quinn, M. S., et al. Explanation in human thinking. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pp. 883–892. PMLR, 2018.

Chi, M. T., De Leeuw, N., Chiu, M.-H., and Lavancher, C. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477, 1994. ISSN 0364-0213. doi: https://doi.org/10.1016/0364-0213(94)90016-7. URL https://www.sciencedirect.com/science/article/pii/0364021394900167.

Crutch, S. J., Connell, S., and Warrington, E. K. The different representational frameworks underpinning abstract and concrete knowledge: Evidence from odd-one-out judgements. *Quarterly Journal of Experimental Psychology*, 62(7):1377–1390, 2009.

Dasgupta, I. and Gershman, S. J. Memory as a computational resource. *Trends in Cognitive Sciences*, 2021.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., and Kurth-Nelson, Z. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

Dove, G. More than a scaffold: Language is a neuroenhancement. *Cognitive Neuropsychology*, 37 (5-6):288–311, 2020. doi: 10.1080/02643294.2019.1637338. URL https://doi.org/10.1080/02643294.2019.1637338. PMID: 31269862.

Edmiston, P. and Lupyan, G. What makes words special? words as unmotivated cues. *Cognition*, 143:93–100, 2015.

Edwards, B. J., Williams, J. J., Gentner, D., and Lombrozo, T. Explanation recruits comparison in a category-learning task. *Cognition*, 185:21–38, 2019.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pp. 1407–1416. PMLR, 2018.

Fodor, J. A. and Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2): 3–71, 1988.

Fyfe, E. R., McNeil, N. M., Son, J. Y., and Goldstone, R. L. Concreteness fading in mathematics and science instruction: A systematic review. *Educational psychology review*, 26(1):9–25, 2014.

Garcez, A. d. and Lamb, L. C. Neurosymbolic ai: the 3rd wave. *arXiv preprint arXiv:2012.05876*, 2020.

Geiger, A., Carstensen, A., Frank, M. C., and Potts, C. Relational reasoning and generalization using non-symbolic neural networks. *arXiv preprint arXiv:2006.07968*, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gentner, D. Why we're so smart. *Language in mind: Advances in the study of language and thought*, 195235, 2003.

Gentner, D. and Christie, S. Relational language supports relational cognition in humans and apes. *Behavioral and Brain Sciences*, 31(2):136–137, 2008.

Ghosh, D., Rahme, J., Kumar, A., Zhang, A., Adams, R. P., and Levine, S. Why generalization in rl is difficult: Epistemic pomdps and implicit partial observability. *Advances in Neural Information Processing Systems*, 34, 2021.

Gopnik, A. and Sobel, D. M. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71(5): 1205–1222, 2000.

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co, 1999.

Goyal, P., Niekum, S., and Mooney, R. J. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020*, 2019.

Gregor, K., Jimenez Rezende, D., Besse, F., Wu, Y., Merzic, H., and van den Oord, A. Shaping belief states with generative environment models for rl. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/2c048d74b3410237704eb7f93a10c9d7-Paper.pdf.

Guan, L., Verma, M., Guo, S., Zhang, R., and Kambhampati, S. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. In *Advances in Neural Information Processing Systems*, 2021.

Hase, P. and Bansal, M. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*, 2021.

Hennigan, T., Cai, T., Norman, T., and Babuschkin, I. Haiku: Sonnet for JAX, 2020. URL http://github.com/deepmind/dm-haiku.

Hermann, K. L. and Lampinen, A. K. What shapes feature representations? exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems*, 2020.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., and Santoro, A. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*, 2019.

Holyoak, K. J. and Lu, H. Emergence of relational reasoning. *Current Opinion in Behavioral Sciences*, 37:118–124, 2021.

Humplik, J., Galashov, A., Hasenclever, L., Ortega, P. A., Teh, Y. W., and Heess, N. Meta reinforcement learning as task inference. *arXiv preprint arXiv:1905.06424*, 2019.

Ichien, N., Liu, Q., Fu, S., Holyoak, K. J., Yuille, A., and Lu, H. Visual analogy: Deep learning versus compositional models. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 2021.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2016.

Jiang, Y., Gu, S., Murphy, K., and Finn, C. Language as an abstraction for hierarchical deep reinforcement learning. *arXiv preprint arXiv:1906.07343*, 2019.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kaplan, R., Sauer, C., and Sosa, A. Beating atari with natural language guided reinforcement learning. *arXiv preprint arXiv:1704.05539*, 2017.

Katz, J. S. and Wright, A. A. Issues in the comparative cognition of same/different abstract-concept learning. *Current Opinion in Behavioral Sciences*, 37:29–34, 2021.

Keil, F. C., Wilson, R. A., and Wilson, R. A. *Explanation and cognition*. MIT press, 2000.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirk, R., Zhang, A., Grefenstette, E., and Rocktäschel, T. A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*, 2021.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Lertvittayakumjorn, P. and Toni, F. Explanation-based human debugging of nlp models: A survey. *arXiv preprint arXiv:2104.15135*, 2021.

Lombrozo, T. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Lombrozo, T. and Carey, S. Functional explanation and the function of explanation. *Cognition*, 99(2):167–204, 2006.

Lombrozo, T. and Vasilyeva, N. Causal explanation. *Oxford handbook of causal reasoning*, pp. 415–432, 2017.

Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., and Rocktäschel, T. A survey of reinforcement learning informed by natural language. *arXiv preprint arXiv:1906.03926*, 2019.

Lupyan, G. Taking symbols for granted? is the discontinuity between human and nonhuman minds the product of external symbol systems? *Behavioral and Brain Sciences*, 31(2):140–141, 2008.

Lupyan, G. The centrality of language in human cognition. *Language Learning*, 66(3):516–553, 2016.

Marcus, G. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

Mu, J., Liang, P., and Goodman, N. Shaping visual representations with language for few-shot classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4823–4830, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.436. URL https://aclanthology.org/2020.acl-main.436.

Mu, J., Zhong, V., Raileanu, R., Jiang, M., Goodman, N., Rocktäschel, T., and Grefenstette, E. Improving intrinsic exploration with language abstractions. *arXiv preprint arXiv:2202.08938*, 2022.

Nam, A. J. H. and McClelland, J. L. What underlies rapid learning and systematic generalization in humans. *arXiv preprint arXiv:2102.02926*, 2021.

Parisotto, E., Song, F., Rae, J., Pascanu, R., Gulcehre, C., Jayakumar, S., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., et al. Stabilizing transformers for reinforcement learning. In *International Conference on Machine Learning*, pp. 7487–7498. PMLR, 2020.

Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*, 2018.

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62 (3):54–60, 2019.

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31 (2):109–130, 2008.

Puebla, G. and Bowers, J. S. Can deep convolutional neural networks learn same-different relations? *bioRxiv preprint*, 2021.

Raileanu, R., Goldstein, M., Yarats, D., Kostrikov, I., and Fergus, R. Automatic data augmentation for generalization in deep reinforcement learning. *arXiv preprint arXiv:2006.12862*, 2020.

Rakelly, K., Zhou, A., Finn, C., Levine, S., and Quillen, D. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Rezende, D. J., Danihelka, I., Papamakarios, G., Ke, N. R., Jiang, R., Weber, T., Gregor, K., Merzic, H., Viola, F., Wang, J., et al. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836*, 2020.

Rittle-Johnson, B. Promoting transfer: Effects of self-explanation and direct instruction. *Child development*, 77 (1):1–15, 2006.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017.

Santoro, A., Hill, F., Barrett, D., Morcos, A., and Lillicrap, T. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pp. 4477–4486, 2018.

Santoro, A., Lampinen, A., Mathewson, K., Lillicrap, T., and Raposo, D. Symbolic behaviour in artificial intelligence. *arXiv preprint arXiv:2102.03406*, 2021.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

Shanahan, M., Nikiforou, K., Creswell, A., Kaplanis, C., Barrett, D., and Garnelo, M. An explicitly relational neural network architecture. In *International Conference on Machine Learning*, pp. 8593–8603. PMLR, 2020.

Sinapov, J. and Stoytchev, A. The odd one out task: Toward an intelligence test for robots. In *2010 IEEE 9th International Conference on Development and Learning*, pp. 126–131. IEEE, 2010.

Stammer, W., Schramowski, P., and Kersting, K. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3619–3629, 2021.

Stephens, R. G. and Navarro, D. J. One of these greebles is not like the others: Semi-supervised models for similarity structures. Proceedings of the 30th Annual Meeting of the Cognitive Science Society, 2008.

Tam, A. C., Rabinowitz, N. C., Lampinen, A. K., Roy, N. A., Chan, S. C., Strouse, D., Wang, J. X., Banino, A., and Hill, F. Semantic exploration from language abstractions and pretrained representations. *arXiv preprint arXiv:2204.05080*, 2022.

Topin, N. and Veloso, M. Generation of policy-level explanations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2514–2521, 2019.

Tulli, S., Wallkötter, S., Paiva, A., Melo, F. S., and Chetouani, M. Learning from explanations and demonstrations: A pilot study. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pp. 61–66, 2020.

Van Fraassen, B. The pragmatic theory of explanation. *Theories of Explanation*, 8:135–155, 1988.

Wang, J. X., King, M., Porcel, N., Kurth-Nelson, Z., Zhu, T., Deck, C., Choy, P., Cassin, M., Reynolds, M., Song, F., et al. Alchemy: A structured task distribution for meta-reinforcement learning. *arXiv preprint arXiv:2102.02926*, 2021.

Watkins, O., Gupta, A., Darrell, T., Abbeel, P., and Andreas, J. Teachable reinforcement learning via advice distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

Williams, J. J. and Lombrozo, T. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive science*, 34(5):776–806, 2010.

Wood, D., Bruner, J. S., and Ross, G. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100, 1976.

Woodward, J. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

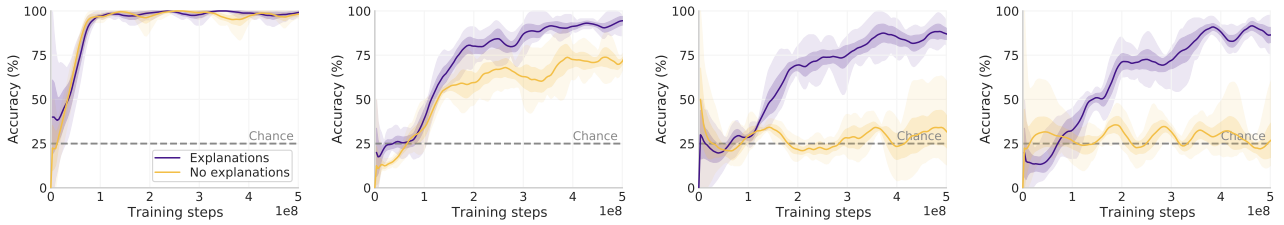
Xie, N., Ras, G., van Gerven, M., and Doran, D. Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*, 2020.

In Appendix A we show additional experiments and analyses. In Appendix B we report numerical values for our main experimental comparisons. In Appendix C we report details of the environments, agents, and training.

A. Ablation experiments & further analyses

In this section, we perform a variety of control, ablation, and auxiliary experiments that identify which attributes of explanations are useful in different settings. We perform most of these experiments in the 2D RL setting because of the efficiency of running and training agents in this environment.

A.1. Agents trained without explanations fixate on the easiest feature dimensions



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

Figure 7: In the 2D setting, agents trained with explanations learn all dimensions, but agents trained without explanations learn to fully solve the tasks only if the relevant dimension is position (the easiest dimension), and only partly learn to solve the tasks with color (the next easiest dimension). (5 seeds per condition.)

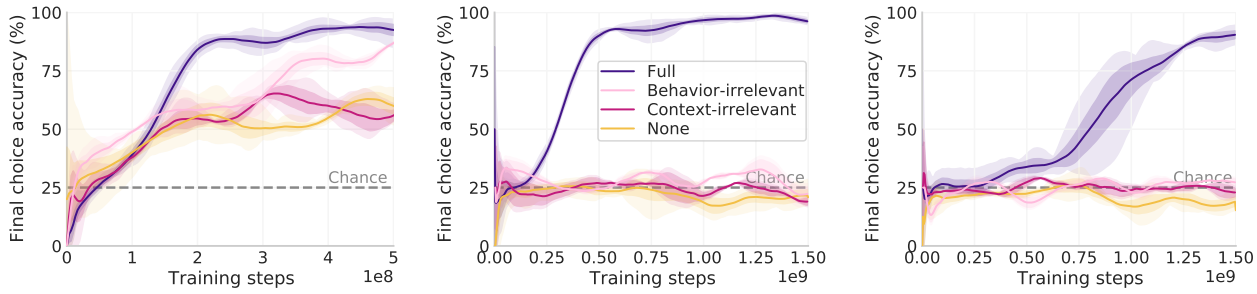
In the basic 2D odd-one-out tasks, the agent achieves off-chance performance without explanations (while in more complicated settings such as the causal interventions, it cannot learn at all without explanations). In Fig. 7, we show that what the agent is doing is latching on to the feature dimension(s) that are most salient and *easiest* (Hermann & Lampinen, 2020), and only correctly solving episodes involving these features. Specifically, position is the most salient feature and is learned rapidly even without explanations, followed by color which is partially learned without explanations. However, shape and texture are much more difficult and are not learned well without explanations. These results concord with the features that Hermann & Lampinen (2020) found were easiest for CNNs and ResNets to learn, suggesting that explanations may help overcome the preference of agents (or other networks) to be "lazy" and prefer "shortcut features" (Geirhos et al., 2020).

A.2. Explanations are most useful if they engage with the agent's behavior; shuffled explanations are useless

We next investigate whether explanations need to be relevant to the agent's behavior, or even to the situation at all, in order to be useful. To do this, we provide the agent with explanations that either are situation-relevant, but behavior irrelevant, or are irrelevant to both behavior and situational context. To produce the situation-relevant but behavior-irrelevant explanations, we first construct an episode as before. We then enumerate all the property and reward explanations that it would be possible to receive in that episode, and present a randomly selected one to the agent on approximately 10% of steps, regardless of the agent's actions. These explanations do contain information about the objects in the scene, and can therefore potentially still benefit learning, but they do not directly react to the agent's actions.

We also considered context-irrelevant explanations that were randomly sampled from the set of all possible explanations (we chose either a property explanation or a post choice one with 50% probability, and then sampled a random set of attributes to fill out the template). This condition is essentially a control for the possibility that predicting structured information—even meaningless information unassociated with the task—could be acting as form of regularization.

Our results (Fig. 8) show that explanations that are relevant to both situation and behavior are most useful, situation-relevant but behavior-irrelevant explanations can be better than nothing in some cases, and totally irrelevant explanations are not beneficial at all. Specifically, for the basic tasks behavior-irrelevant explanations still result in some learning, but are much slower than full behavior-relevant explanations.



(a) Basic 2D. (b) Meta-learning (easy levels). (c) Meta-learning (hard levels).

Figure 8: Explanations must be behaviourally (as well as contextually) relevant to be useful in challenging settings; explanations that are contextually-irrelevant are useless in every experiment. (a) In the basic 2D odd-one-out tasks, behavior-irrelevant explanations eventually result in relatively comparable performance compared to full explanations, but produce much slower learning. Context irrelevant explanations are not substantially different than no explanations. (b-c) In both easy and hard learning-to-experiment levels, only an agent with full, behavior and context-relevant explanations is able to learn the tasks at all. Thus, more challenging task settings require more specific, behavior-releavnt explanations. (3 seeds for behavior-irrelevant/context-irrelevant conditions.)

A.3. Providing explanations as agent inputs is not beneficial, and interferes with learning from explanation targets

In the main text, we focused on explanations as targets during training, rather than inputs. That approach is beneficial, because it does not require explanations at test time, while explanations as input generally does. In Fig. 9, we show that furthermore *providing explanations as input to the agent is not beneficial, and is actively detrimental if explanations are also used as targets*, presumably because in the latter case the agent can just "pass through" the explanations, without having to learn the task structure. However, it is possible that providing explanations as input on the timestep *after* the agent predicts them could be useful (as in language model prediction, where each word is input after the model predicts it); we leave evaluating this possibility to future work.

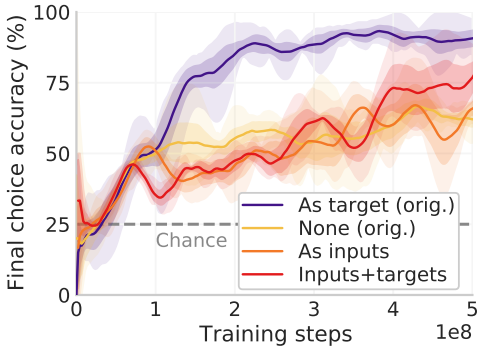


Figure 9: Providing explanations as agent inputs is not beneficial (performs better than no explanations), and is actively detrimental if explanations are also used as targets. (3 seeds per as-input condition, 5 seeds for main conditions.)

A.4. Different kinds of explanations have complementary, sometimes separable benefits

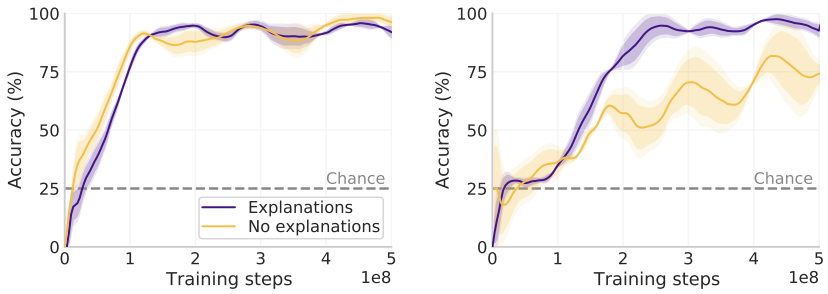
We generally provided agents with both property explanations and reward explanations. Is one of these explanations more useful than the other? Are they redundant? To answer these questions, we considered providing the agent with each kind of explanation independently. We generally find that having both types of explanations is best, and the benefits of different types depend on the setting.

In the 2D setting (Fig. 6a) either type of explanations alone results in learning, but both types together result in substantially

faster learning. In the 3D setting (Fig. 6b), we find that property explanations are uniquely beneficial; perhaps because predicting explanations on encountering the objects helps the agent overcome the challenges of generating good representations for its memory.

For the meta-interventions tasks involving experimentations, we find (Fig. 6c) that the reward explanations are uniquely beneficial, while properties explanations are not useful alone. The likely reason for this is clear when considering the episode structure—the relevance of a transformation to the final reward is much more directly conveyed by the reward explanations than the property ones. However, both types of explanations together are required for complete learning within the learning time we considered.

A.5. The benefits of explanations depend on task complexity



(a) Position & color only.

(b) Shape & texture only.

Figure 10: The benefits of explanations depend on task difficulty. We train agents in the 2D environment on easier odd-one-out tasks where only two of the four dimensions are ever relevant—in these easier cases the agent is able to make more progress without explanations. (a) When only the easy dimensions of position and color are ever relevant, the agents trained without explanations learn just as rapidly as the agents trained with explanations. (b) When the agents are trained on levels where only the harder dimensions of shape and texture are relevant, explanations still accelerate learning substantially. However, the agents trained without explanations achieve some learning in this condition, while they do not achieve any learning on these dimensions in the harder tasks used for the main experiments (see Fig. 7c-d). (2 seeds per condition.)

While we generally considered tasks with many feature dimensions that might be relevant, here we show that simpler tasks in which only two dimensions vary do not always require explanations for learning. However, task complexity depends on both the number of possibly relevant dimensions and the base difficulty of those dimensions. In Fig. 10 we show that explanations are not beneficial compared to no-explanations when only the easy features of position and color are relevant. Explanations are still beneficial when the features are more difficult (shape and texture). But even in this condition, the agent without explanations exhibits some learning, while it does not learn these dimensions at all in the main experiments, where the easier dimensions are also included (see Fig. 7c-d). Note also our results on deconfounding—in some cases explanations may help the agent to generalize in a desired way even if they are not necessary for learning the training task.

A.6. Language prediction is learned faster than RL

In Fig. 11 we show that the language loss decreases substantially early in training, before the agent has mastered the tasks—although the reward explanations are learned more slowly than property explanations, both are substantially learned before the agent masters the tasks. This supports our hypothesis that language, by providing a more consistent signal that emphasizes the task structure, makes this structure more learnable, and thereby supports RL.

A.7. Learning properties through a curriculum rather than auxiliary losses

Because predicting property explanations alone can be beneficial, we next consider whether the agent could benefit from learning properties through auxiliary tasks which teach those properties, rather than through explanations. Specifically, we provide the agent with a simpler property-learning task in 50% of episodes, where it receives a property like “red” as an input instruction, and has to choose the corresponding object (all objects are different along each feature dimension). These

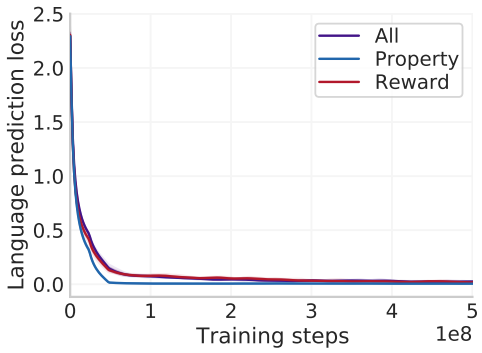


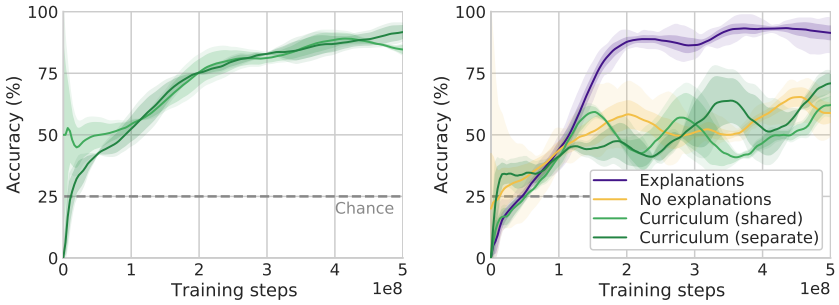
Figure 11: The language prediction loss decreases rapidly early in training, relative to when the agent learns the RL tasks (evaluated on the basic 2D tasks with either full explanations, or just property or reward; compare to RL learning dynamics in Fig. 6a). The agent trained only with property explanations learns to predict them especially rapidly—predicting reward explanations requires more learning, unsurprisingly because these require mastering more of the task structure. However, in all cases the language prediction loss decreases substantially before the agent masters the task. (5 seeds for all explanations condition, 2 for others. Note that the lines on this plot represent different training conditions that include all or some of the explanations; we were unable to separate the language loss for the different types of predictions for the agent trained with both types.)

tasks provide a different way to force the agent to learn the properties of the objects. On the odd-one-out tasks, the agent receives the instruction "find the odd one out" to distinguish its goal.

Surprisingly, we find that learning these tasks does not substantially accelerate learning of the odd-one-out tasks compared to a no-explanations and no-property-tasks baseline (Fig. 12b). We initially thought this might be due to interference due to the shared policy being used for different tasks, so we reran the experiment with separate policy heads for the curriculum tasks and odd-one-out tasks, but this did not substantially change results. Auxiliary prediction of explanations may therefore be a more efficient way to encourage learning of task-relevant features, perhaps because it actively engages with the agent's behavior in the settings where those dimensions are particularly relevant.

A.8. Auxiliary unsupervised losses are neither necessary nor sufficient; thus the benefits of explanations are not simply due to more supervision

In Fig. 13 we show that the auxiliary reconstruction losses are not necessary for learning the odd-one-out tasks. Furthermore, the main text results without explanations show that these losses are not sufficient for learning either. This shows that the benefits of explanations are not simply due to having more supervision for the agent, but rather are specific to supervision that highlights the abstract task structure.



(a) Property-curriculum tasks.

(b) Odd-one-out tasks.

Figure 12: Agents trained with a curriculum of tasks that teach the properties (a) do not learn the odd-one-out tasks (b) any better than agents trained with no explanations. Results are similar whether the agent uses a shared policy (light green) for both the curriculum and odd-one-out tasks, or uses separate policies for each (dark green). (2 seeds per condition for curriculum conditions, 5 seeds for main conditions.)

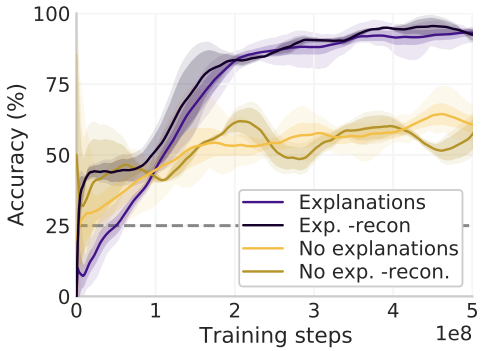


Figure 13: Ablating the auxiliary reconstruction losses does not alter the pattern of results in the 2D environment—thus reconstruction is not necessary for learning these tasks. Comparing to the results which include the reconstruction losses shows that reconstruction losses are also not sufficient without explanations, even though the reconstruction losses provide substantially more supervision on every step. (2 seeds per condition.)

B. Quantitative results

Table 1: Numerical results from main experiments/figures in each domain—mean ± standard deviation across seeds. Results are average performance (% correct) across evaluations during the last 1% of training.

Experiment	Level	Fig.	Condition	Performance
Perceptual 2D	-	3b	Explanations	91.3 ± 0.7
			No explanations	61.9 ± 2.2
Perceptual 3D	-	3c	Explanations	92.7 ± 1.4
			No explanations	29.5 ± 0.7
Deconfounding	Chose color	4b	No explanations	55.4 ± 2.6
	Chose shape		No explanations	24.2 ± 7.6
	Chose texture		No explanations	15.4 ± 6.7
	Chose color	4c	Explain color	95.5 ± 0.9
	Chose shape		Explain shape	87.5 ± 2.9
	Chose texture		Explain texture	86.2 ± 0.9
Meta-learning	Easy	5b	Explanations	96.9 ± 0.3
			No explanations	24.0 ± 0.6
	Hard	5c	Explanations	90.5 ± 1.6
			No explanations	24.6 ± 1.2

C. Methods

C.1. RL agents & training

Table 2: Hyperparameters used in main experiments. Where only one value is listed across both columns, it applies to both.

	2D	**3D**
All activation fns	ReLU	
State dimension	512	
Memory dimension	512	
Memory layers	4	
Memory num. heads	8	
TrXL extra length	128	
Visual encoder	CNN	ResNet
Vis. enc. channels	(16, 32, 32)	
Vis. enc. filt. size	(9, 3, 3)	(3, 3, 3)
Vis. enc. filt. stride	(9, 1, 1)	(2, 2, 2)
Vis. enc. num. blocks	NA	(2, 2, 2)
Policy & value nets	MLP with 1 hidden layer with 512 units.	
Reconstruction decoder	Architectural transpose of the encoder, with independent weights.	
Explanation decoder	1-layer LSTM	
Explanation LSTM dimension	256	
Recon. loss weight	1.	0.
V-trace loss weight	1.	
V-trace baseline weight	0.5	
Explanation loss weight	$3.3 \cdot 10^{-2}$ for main, 0.2 for meta-learning.	$3.3 \cdot 10^{-2}$
Entropy weight	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$
Batch size	24	
Training trajectory length	50	
Optimizer	Adam (Kingma & Ba, 2014)	
LR	$1 \cdot 10^{-4}$	$5 \cdot 10^{-5}$ with explanations, $2 \cdot 10^{-5}$ without

In Table 2 we list the architectural and hyperparameters used for the main experiments. All agents were implemented using JAX (Bradbury et al., 2018) and Haiku (Hennigan et al., 2020), and were trained using TPU v3 and v4 devices.

Additional architectural details: The agent visual encoder output is flattened and then linearly projected to a single vector; the transformer memory receives this one vector input per timestep, and the corresponding single transformer memory output vector for the current timestep is used as input to the heads. The transformer memory uses relative position encodings. The visual decoders use depth-to-space upsampling where necessary, i.e. they create a feature map of shape $N \times M \times (C \cdot k^2)$ and then reshape by location to get a result of shape $(N \cdot k) \times (M \cdot k) \times C$. The agent outputs a full language prediction per timestep

Hyperparameter choices and tuning: In most cases the hyperparameters were taken from other sources without tuning for our setup. There are two main exceptions: (1) we chose the explanation weight loss to approximately balance the magnitude of this loss with the magnitudes of the RL losses early in training in each experiment, and (2) we swept the learning rate for the main experiments in 2D and 3D (but used the same settings for follow-up experiments, including meta-learning). In the 2D experiments we found that a similar learning rate was best for both agents trained with and without explanations, but in 3D we found that agents trained without explanations needed a slower learning rate to avoid their performance degrading from chance-level to below chance.

Since we ran the 3D experiments after 2D, we used similar hyperparameters, except that we found we needed to decrease the learning rate as noted above. However, some hyperparameters do differ across tasks due to specific task features. For example, the visual encoder for the 2D tasks is set to have a filter size of 9 because this is the resolution of each square in the grid, and the entropy cost for 2D tasks was chosen from prior work which used a similar grid world action space (Hill et al.,

2019), while the 3D cost is lower because of the more complex action space for these tasks (see below). These decisions were shared across experimental conditions, and the choices were based upon prior work in similar environments, so should not favor one condition over another.

Explanation prediction loss: We trained the agents to predict the language explanation using a softmax cross-entropy loss over a word-level vocabulary of 1000 tokens (more than were necessary for the limited language we used). The explanation loss was summed across the sequence of tokens.

Self-supervised image reconstruction loss: We trained the agents to reconstruct the image pixels (normalized to range [0, 1] on each color channel) with a sigmoid cross-entropy loss. The image reconstruction loss was averaged across all pixels and channels. However, we found in follow-up experiments that this did not substantially change results in 2D (Appx. A.8), so we disabled this loss in 3D.

C.2. RL environment details

We are in the process of preparing our 2D environments for release; once this process is complete they will be released at https://github.com/deepmind/tell_me_why_explanations_rl

C.2.1. 2D

The 2D tasks were implemented in Pycolab (https://github.com/deepmind/pycolab). instantiated in a 9×9 tile room with an extra 1 tile wall surrounding on all sides, for a total of 11×11 tiles. This was upsampled at a resolution of 9 pixels per tile to form a 99×99 image as input to the agent. The agent was placed in the center of the room, and had 9 possible actions, allowing it to move one square in any of the 8 possible directions, or to do nothing.

Four objects were placed in the room with the agent. They were chosen so that a single object was the odd one out, along a single dimension, and features appeared in two pairs along the other dimensions. The objects varied along the feature dimensions of:

- Color: one of 19 possible colors (e.g. green or lavender).
- Shape: one of 11 possible shapes (e.g. triangle or tee).
- Texture: one of 6 possible textures (e.g. horizontal stripes or checkers).
- Position: One of 4 position types (in corner, against horizontal wall, against vertical wall, or in a 3x3 square in the center).

The agent was given 128 steps to complete each episode, after which the episode would immediately terminate. The agent had to choose an object by walking onto the grid cell containing it. It would be immediately rewarded 1 if the object was the odd one out, and 0 otherwise. However, the episode would last for an additional few steps to give the agent time to learn from the reward explanation (if provided); this extra time was provided even if the agent was not trained to predict explanations, in order to precisely match the training experience across conditions. No additional reward would be received during this period, but the agent would be asked to output the reward explanation at every timestep. This period would last either 16 steps, until the agent touched any object, or until the full episode limit of 128 steps was reached, whichever came soonest. (However, this full extra length was likely unnecessary, see below.)

If the agent was trained to predict property explanations, whenever it was adjacent to an object it would be asked to predict a string of the form:

```
This is a red horizontal-striped triangle in-the-corner
```

The properties always appeared in the order color texture shape position. These sentences were tokenized at a word level, with the hyphenated phrases treated as single words. Hence, a single token was attached to each possible feature value along each dimension.

If the agent was adjacent to multiple objects, which description it received was determined randomly. Once the agent made a choice, property descriptions were disabled.

If the agent was asked to predict reward explanations, for the period after receiving the reward (see above), it would be asked to predict a string in one of the following forms:

```
Correct because it is uniquely horizontal-striped

Incorrect because other objects are red horizontal-striped
triangles or in-the-corner
```

Thus, the reward explanations identify all features that contributed to a decision being incorrect.

Does the agent need to predict the reward explanation for 16 steps? We gave the agent 16 steps to predict the reward explanations to make the signal more salient. However, in a follow-up experiment (not shown) where the agent had only a single step for prediction in 2D domain, we observed only a very minor slowdown in learning, even if the agent was only given reward explanations. Thus we do not believe that this extra time is strictly necessary (although we have not thoroughly explored this across all experimental domains).

Confounding: The confounded training tasks were performed in a variation of the 2D setup. All objects were initialized to have the same position type (as defined above; randomly selected), but distinct positions. One object was chosen to be the target. The other three objects were all assigned matching colors, shapes, and textures, while the target object was assigned a unique color, shape, and texture.

When explanations were provided, both property and reward explanations described only a single feature dimension in any given condition.

Deconfounded tests: The deconfounded tests, used for both the confounding experiments and the final trial of the meta-learning experiments (below), assigned all objects to the same position type. A common color, shape, and texture were chosen, as well as rare versions of each. One object was assigned common versions of all three attributes, the second was assigned the rare color but common shape and texture, the third was assigned the rare shape but common color and texture, and the final one was assigned the rare texture but common shape and color.

Meta-learning to experiment: For the experimentation tasks, the agent received a series of four trials. The first three trials were experimentation trials, while the fourth was a deconfounded test (as above), except that on the final trial the agent received a reward of 10 for a correct answer instead of 1. The agent was given 512 steps to complete all 4 trials—i.e., average time allotted per trial equal to the tasks above.

For the experimentation trials, the agent was granted three additional actions, which would transform the corresponding attributes (color, shape, or texture) of an adjacent object. If the agent was adjacent to multiple objects, the target was chosen pseudo-randomly (according to the order in the environment's internal object representation). The agent could only employ one transformation action per trial. After the first time, these actions did not have any effect. If the agent was not adjacent to an object, or during the deconfounded test trial, the actions also had no effect.

The objects for the experimentation trials were again initialized with matching position types. The other attributes could either all be matching (easiest levels), or could be paired (harder levels)—in the latter case, there would be two values for that attribute that each appeared on two objects, e.g., two red objects and two blue. The results shown in the main figures are from levels where all are matching, or all are paired, but we also trained the agent on intermediate levels where some attributes were paired, but some were matching, which helped the agent to transition from learning the easy levels to the hardest ones.

The reward explanations that agents received named which dimension was relevant in the current episode. Neither property descriptions nor reward explanations mentioned position, as it was never a relevant concept.

In order for the agent to determine which type of trial it was performing, it was allowed to observe an input instruction that either said "Make an odd one out" (experimentation trials) or "Find the odd one out" (final trial). This was provided through a separate channel from the explanations; as before, the agent did not observe the explanations.

C.2.2. 3D

The 3D environments were implemented in Unity. The agent was placed in a room with a randomly-located door and windows (the agent could not interact with these). It had 10 possible actions: moving forward or backard, moving left or right, looking left or right or up or down, grabbing an object it was facing (if within a certain distance), and doing nothing.

Four objects were placed in the room, with attributes sampled as in the 2D environment from the dimensions:

- Color: one of 10 possible colors (e.g. blue or magenta).
- Size: one of 3 possible sizes (small, medium, or large).
- Texture: one of 6 possible textures/materials (e.g. metallic or wood-grain).
- Position: One of 3 position types (in corner, against wall, or in the center).

The episode lasted for 60 seconds, at 30 FPS; but the agent took an action only once every 4 frames (the action was then repeated until the next agent step), so the episode lasted for at most 450 agent steps. The agent had to use its "grab" action on an object to make a choice; colliding with the objects would simply cause them to move. 25 steps were allocated for reward explanations (if any).

If the agent was asked to predict property explanations, they were given when the agent was facing an object and close enough to grab it.

The property and reward explanations for the 3D environment were analogous to the ones for the 2D environment, except that the reward explanations did not have the "or" before the final attribute.