Goal Misgeneralization as Implicit Goal Conditioning

Diego Dorn
EPFL
diego.dorn@epfl.ch

Neel Alex
University of Cambridge

David Krueger
University of Cambridge

Abstract

While many examples of goal misspecification [2] have been dissected in the reinforcement learning literature, few works have focused on the relatively new goal misgeneralization [3, 5]. As goal misgeneralization often stems from underspecification, we explore a simple environment with some goals specifiable through explicit conditioning, and others not. We find that agents generally pursue a mixture of possible goals, and the choice of goal to pursue is often inexplicable. Nonetheless, we attempt an explanation of *implicit goal conditioning* – wherein subtle environment features determine which goal is pursued – and aim to understand which features induce pursuit of one goal over another.

1 Introduction

Langosco et al. [3] identify a few simple examples of goal misgeneralization. For instance, an agent trained in ProcGen's CoinRun environment [1] with the coin's position fixed to the rightmost side of the screen typically learns to run to the right rather than collect coins, when evaluated in environments where the coin's position is randomized. While the agent is highly capable – navigating obstacles and avoiding enemies, it will run up to the right wall and stay there unless the coin location experiences enough randomization in training.

This behavior arises because the training environment does not disambiguate between possible goals. We consider this behavior worthy of study, as goal specifications and training environments will always be ambiguous, and the difficulty of doing so is highest in complex, real-world environments. In such situations, it is particularly important to understand and predict how learned policies generalize.

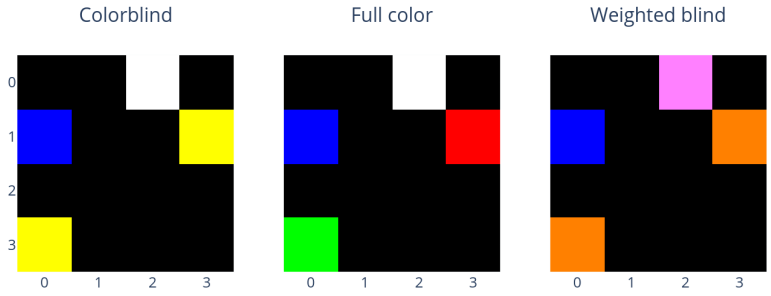

Figure 1: Left: The agent's training environment, in which the red and green goals are ambiguous. Center: The agent's testing environment, in which the red and green goals are no longer ambiguous. Right: An alternate training environment, in which additionally, signal on the green channel is weaker. In each setup, the agent also receives a one hot encoded input representing the target goal.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

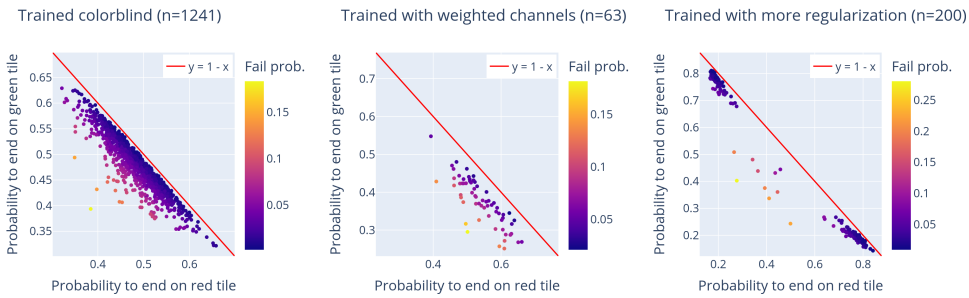

Figure 2: We train multiple agents in each environment, then evaluate 10,000 times in the full-color environment to determine its probability of going to red vs. green. Left: Trained in colorblind. Center: Trained in weighted colorblind. Right: Trained in colorblind with lowest-channel regularization.

2 Experiments

To elicit this behavior, we create [1] a simple 4x4 gridworld environment with three goal types: red, blue, and green, visible in Figure 1. The agent is conditioned on a specific goal type. However, to create goal ambiguity, the agent cannot distinguish the red and green goals. We call this training mode "colorblind", and set the red and green channels in the agent's observation to the maximum of the red and green channels (including the goal conditioning). We also create a weighted colorblind mode, identical to the colorblind mode except that after calculating the maximum between channels, the green channel is then multiplied by 0.5. In testing, the agent is always able to see in full color.

We train deep reinforcement learning agents with PPO [4], providing goal conditioning as a part of the agent's observations. Our initial results are visible in Figure 2. When trained in the typical colorblind setting, agents do not exhibit any single generalization behavior. Instead, any individual agent may choose to follow the red or the green goal with some probability. Curiously, this probability is not fixed by the parameters of the training set-up, but has substantial variance between agents!

We demonstrate that modifications to the training set-up can alter the induced distribution of agents. In the weighted colorblind environment, we see that the distribution favors red. Additionally, adding a regularization penalty to the channel in the agent's first layer with the lowest ℓ_2-norm produces a bimodal distribution of agents, where individual agents are likely to favor one goal over the other.

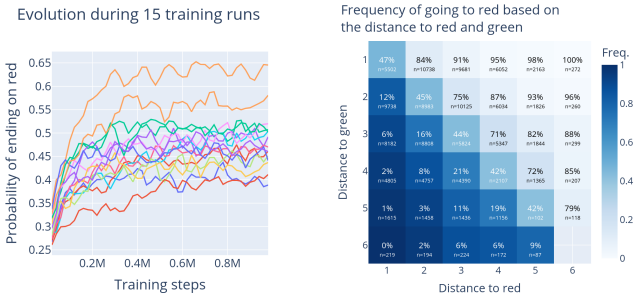

Figure 3: Left: fraction pursuit of red goal for many agents over the course of training. Right: fraction pursuit of red goal for a single agent as a function of distance to each goal.

3 Conclusion

As Figure 3 shows, this behavior can be attributed to neither unconverged training, nor wholly to the distance to the nearest goal! Rather, training runs with different random seeds and individual agents in different environments can have surprising variance in final behavior. More investigation is required

[1] Code is availaible at `https://github.com/ddorn/seed-dependent-goals`

to determine what effect this has on more complex agents, and to what extent this generalization can be controlled so that agents can more readily follow the designer's intended goals in deployment.

References

[1] Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. *arXiv preprint arXiv:1912.01588*, 2019.

[2] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of ai ingenuity, 2020. URL `https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity`.

[3] Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. Goal misgeneralization in deep reinforcement learning, 2023.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[5] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals, 2022.