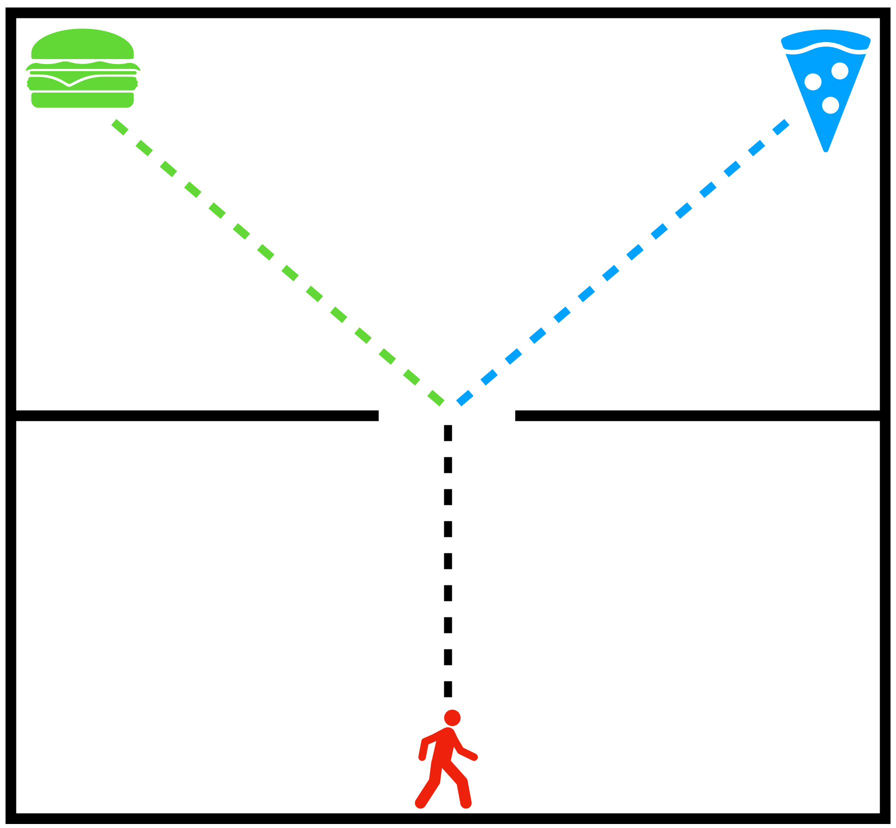
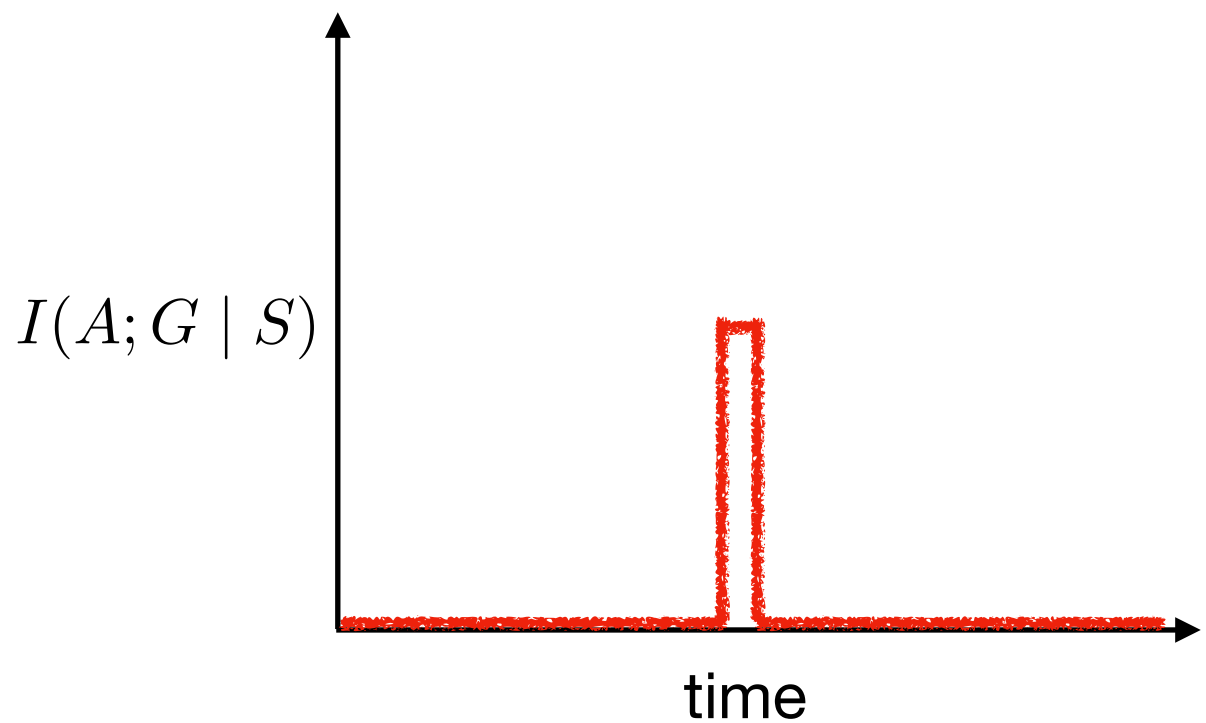
Hierarchical Reinforcement Learning via Information Bottleneck

DJ Strouse
with Jane Wang, David Pfau, Neil Rabinowitz, & Matt Botvinick

TaCL, October 3, 2017

Information and subgoals



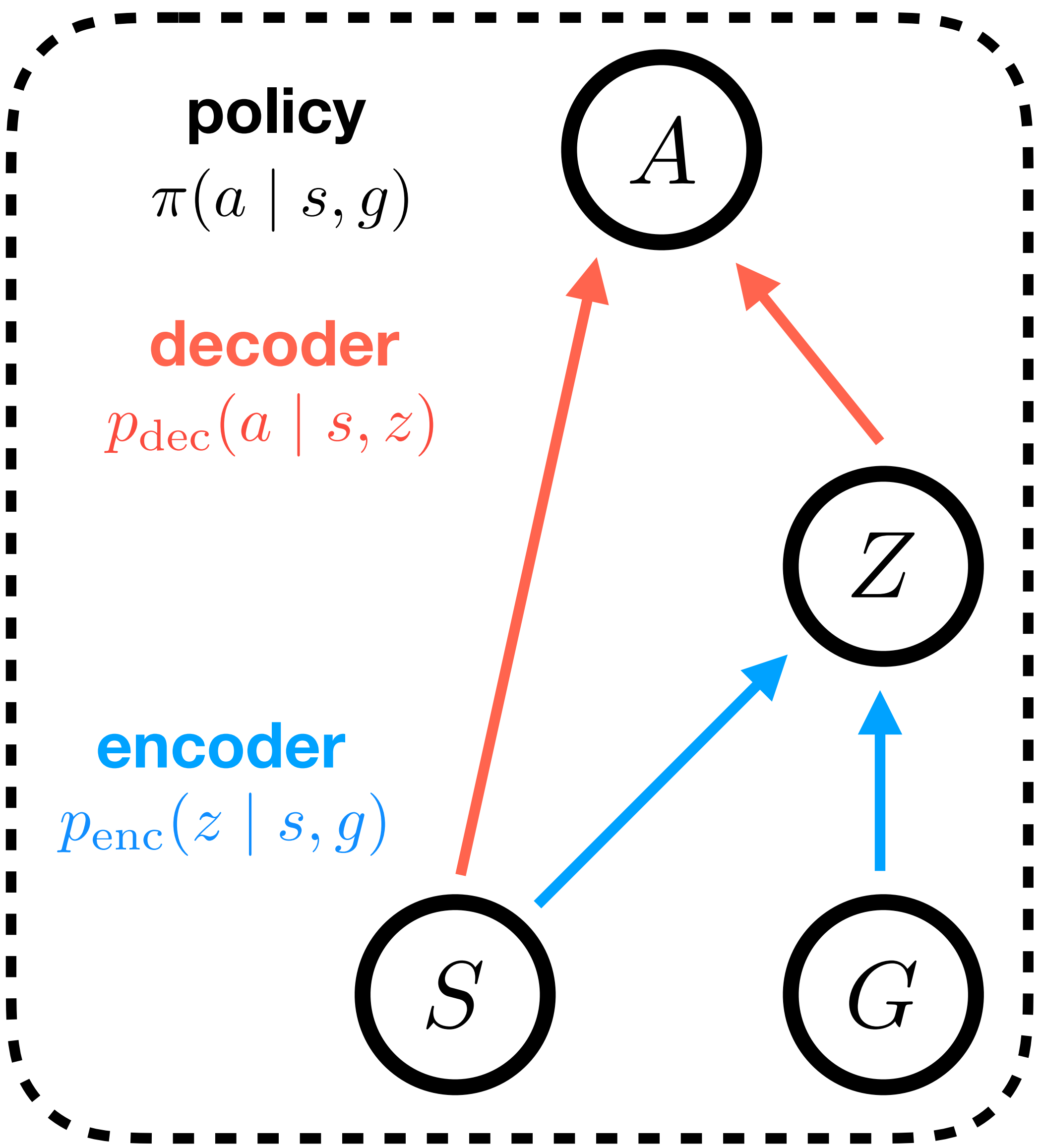


Information identifies useful subgoals

van Dijk & Polani, Grounding Subgoals in Information Transitions, 2011

Information regularizer -> encourages efficient hierarchical policies?

Regularization via goal bottleneck



variational information minimization

$$I(A; G \mid S) \leq I(Z; G \mid S)$$

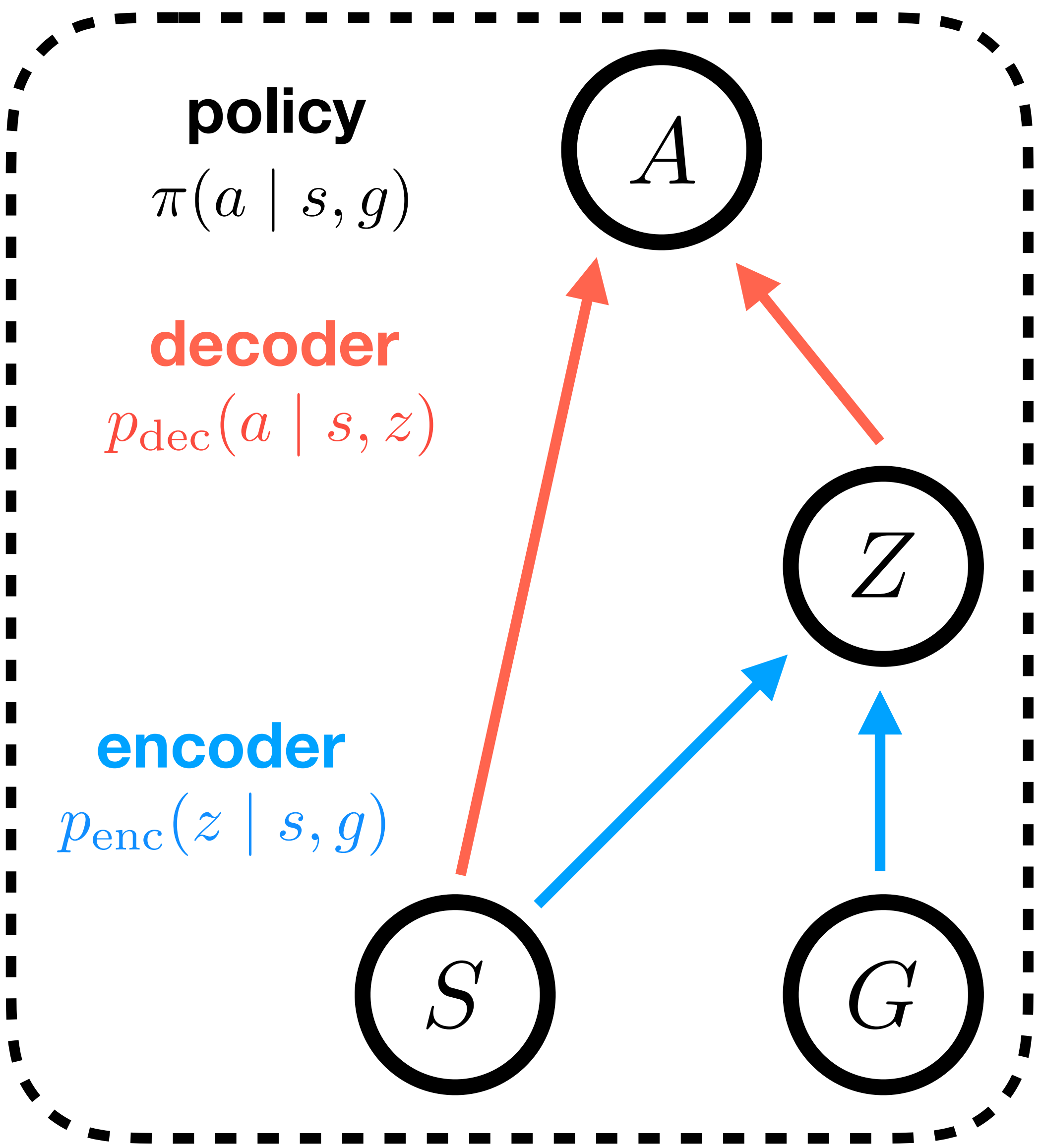
$$\leq \sum_g p(g) \sum_s p(s \mid g)\, \mathrm{KL}[p_{\text{enc}}(z \mid s, g) \mid r(z)]$$

sample a goal — *sample trajectory* — *penalize encoder for departures from prior*

interpretations

- communication bottleneck between goal & agent
- encourage “habits”
- minimize cognitive cost of control
- reduce load on working memory
- (lossy) policy compression: $\pi(a \mid s, g) \approx \pi(a \mid s)$
- decoder should develop a language of relevant behaviors, which the encoder learns to speak in

Regularization via goal bottleneck



variational information minimization

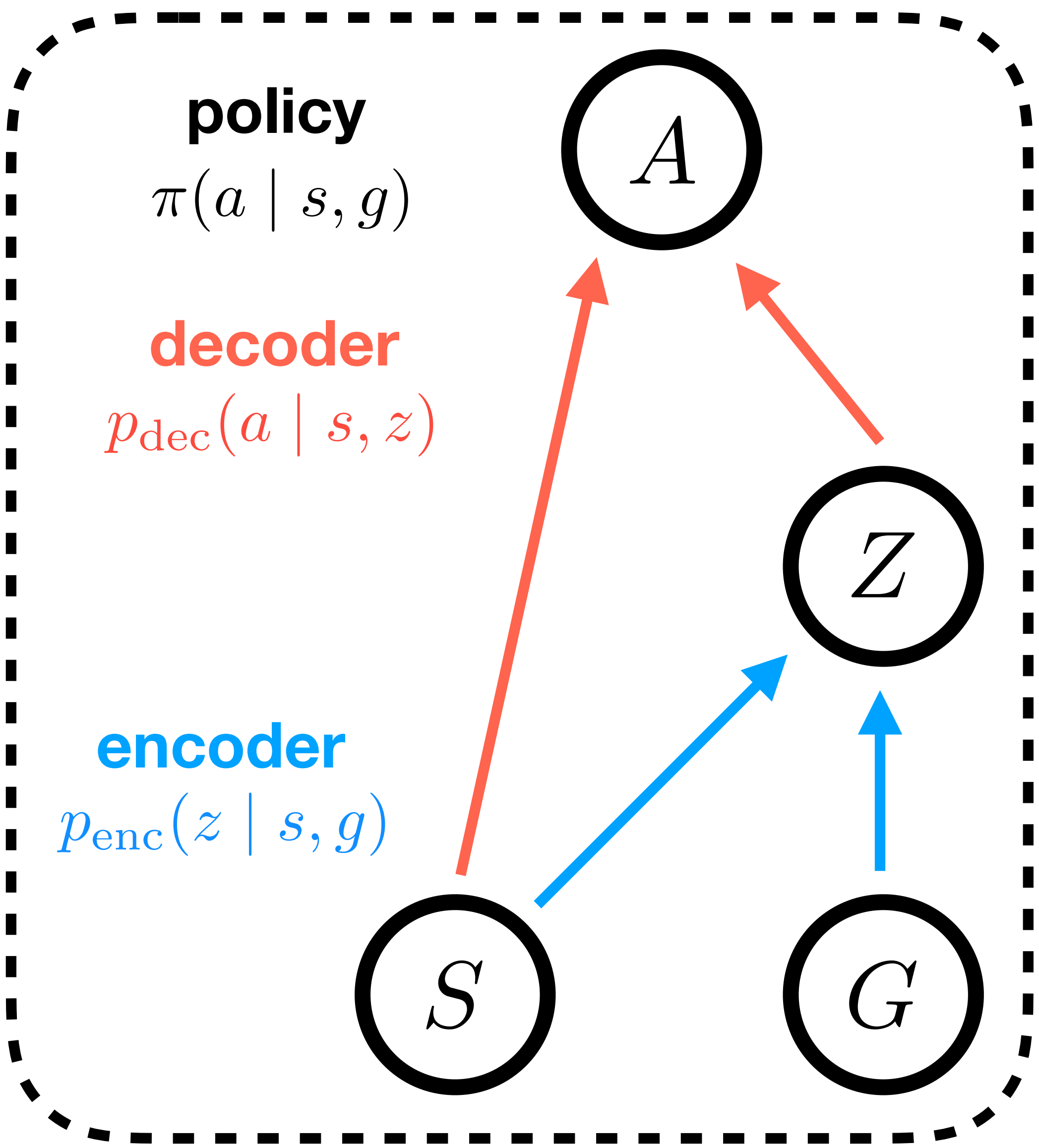
$$I(A; G \mid S) \leq I(Z; G \mid S)$$
$$\leq \sum_g p(g) \sum_s p(s \mid g)\, \mathrm{KL}[p_{\text{enc}}(z \mid s, g) \mid r(z)]$$

sample a goal — *sample trajectory* — *penalize encoder for departures from prior*

additional details

- train using REINFORCE with state-value baseline
- 2 regularizations: above + entropy
- tabular encoder / decoder
- discrete latents (marginalized out; not sampled)
- fixed uniform prior

Related work: VIB



variational information minimization

$$I(A; G \mid S) \leq I(Z; G \mid S)$$
$$\leq \sum_g p(g) \sum_s p(s \mid g)\, \mathrm{KL}[p_{\text{enc}}(z \mid s, g) \mid r(z)]$$

sample a goal — *sample trajectory* — *penalize encoder for departures from prior*

variational information bottleneck (VIB)
Alemi, Fischer, Dillon, & Murphy 2017

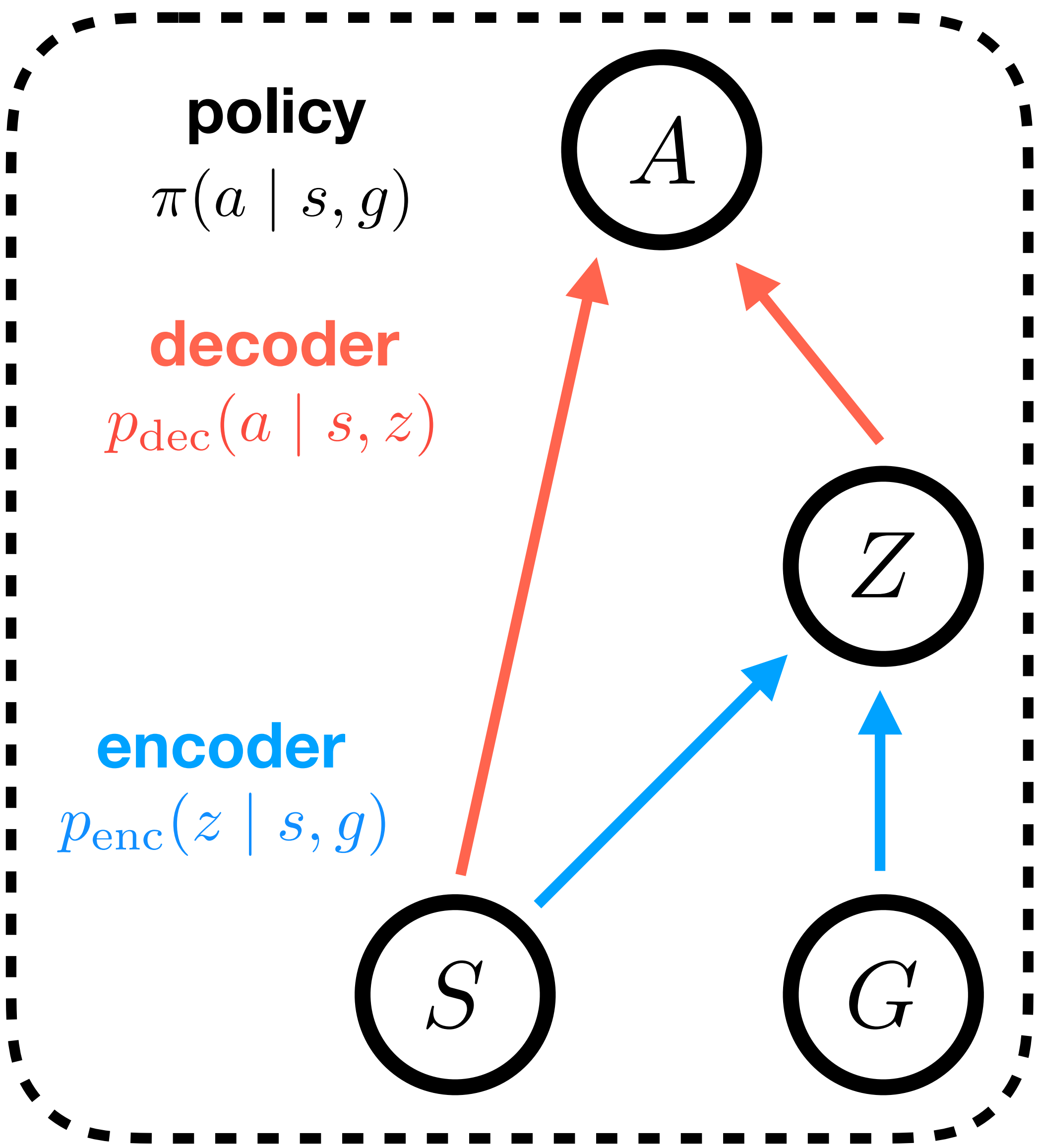
difference of mutual informations

$$L = I(Z; G \mid S) - \beta I(A^*; Z \mid S)$$ *correct action*

$$\leq \sum_g p(g) \sum_s p(s \mid g)\, [-\log \pi(a^* \mid s, g) + \beta \cdot \mathrm{KL}]$$

-> regularized imitation learning

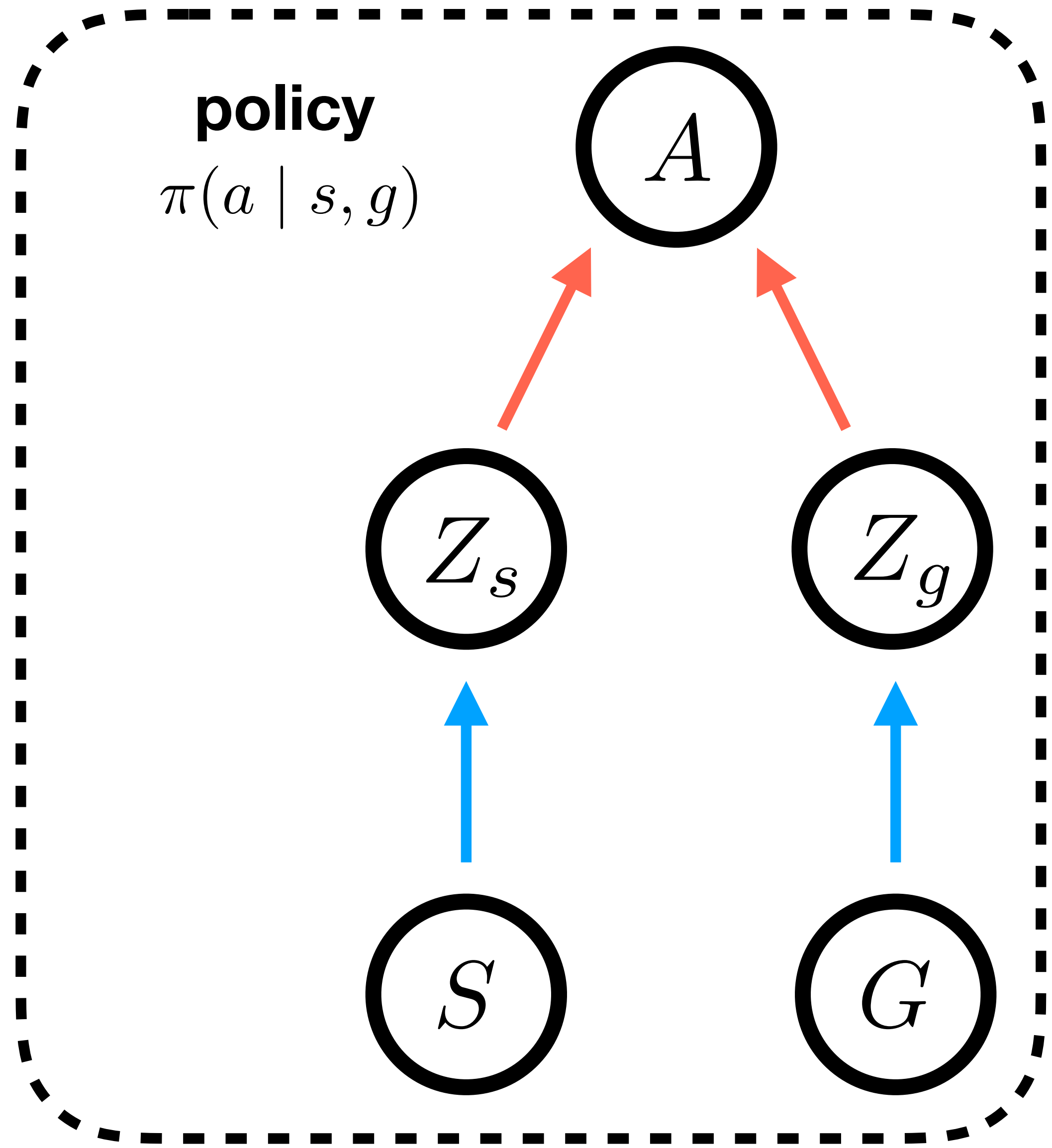
Related work: UVFA



info bottleneck on latents

universal value function approximator (UVFA)

Schaul, Horgan, Gregor & Silver 2015



physical bottleneck on latents

Related work: info regularizers

- State bottleneck (Tishby, Polani, & others)

$$\min \ I(S_{\text{current}}; A)$$

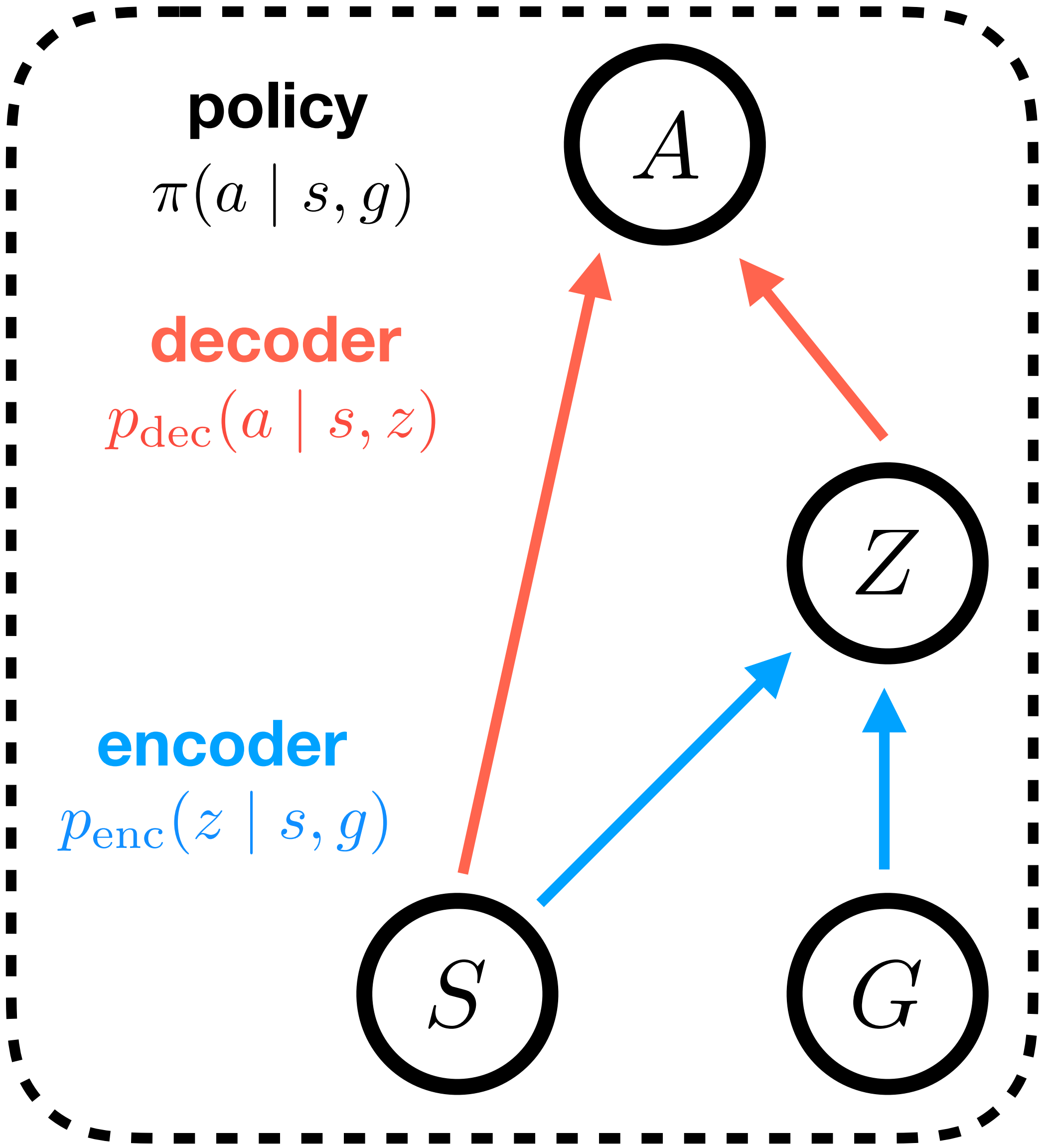
- Empowerment (Polani, Mohamed, Rezende, & others)

$$\max \ I(A; S_{\text{future}})$$

- Variational intrinsic control (Gregor, Rezende, Wierstra)

$$\max \ I(\text{set of options}; \text{option termination states})$$

Regularization via goal bottleneck



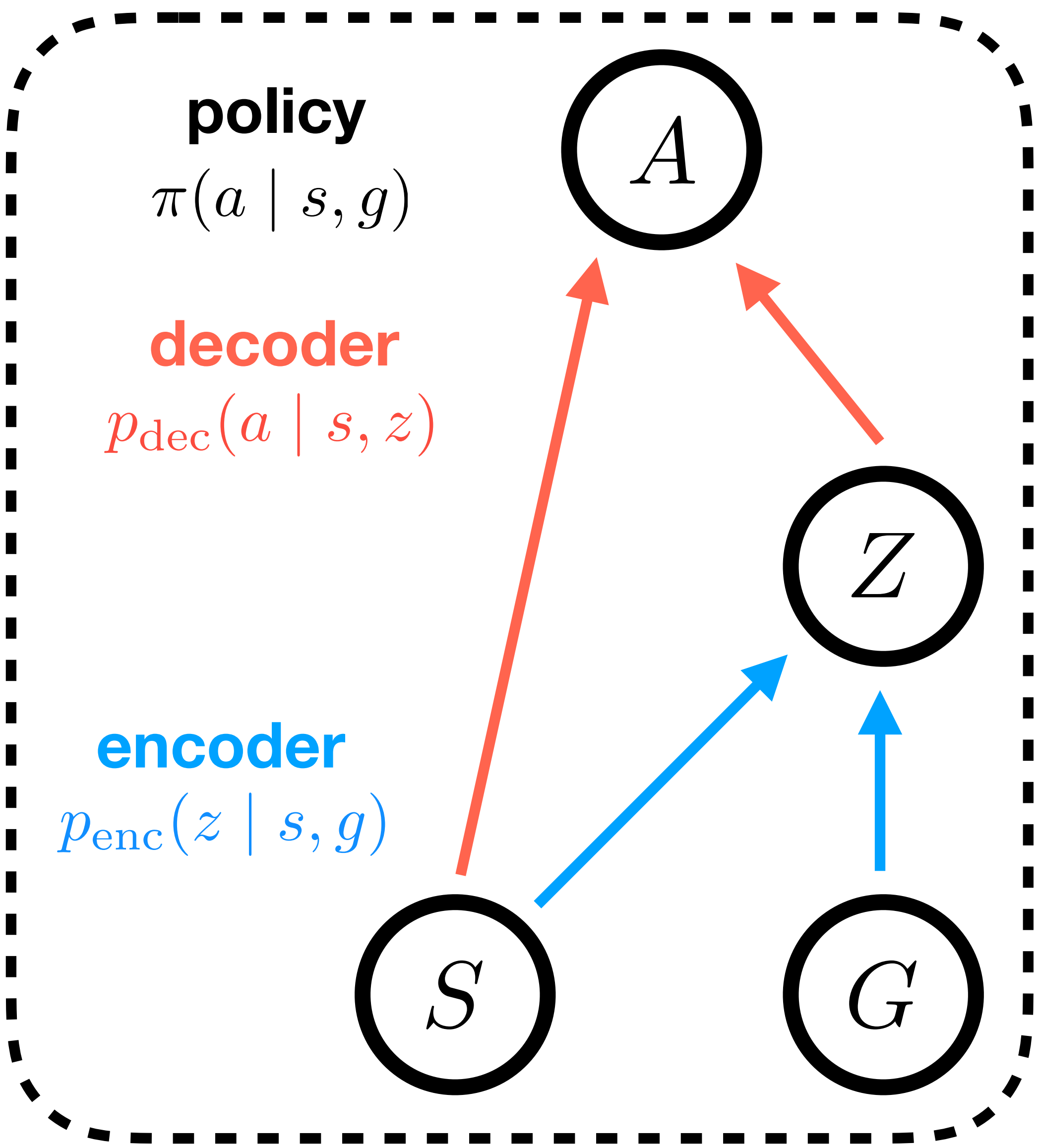
variational information minimization

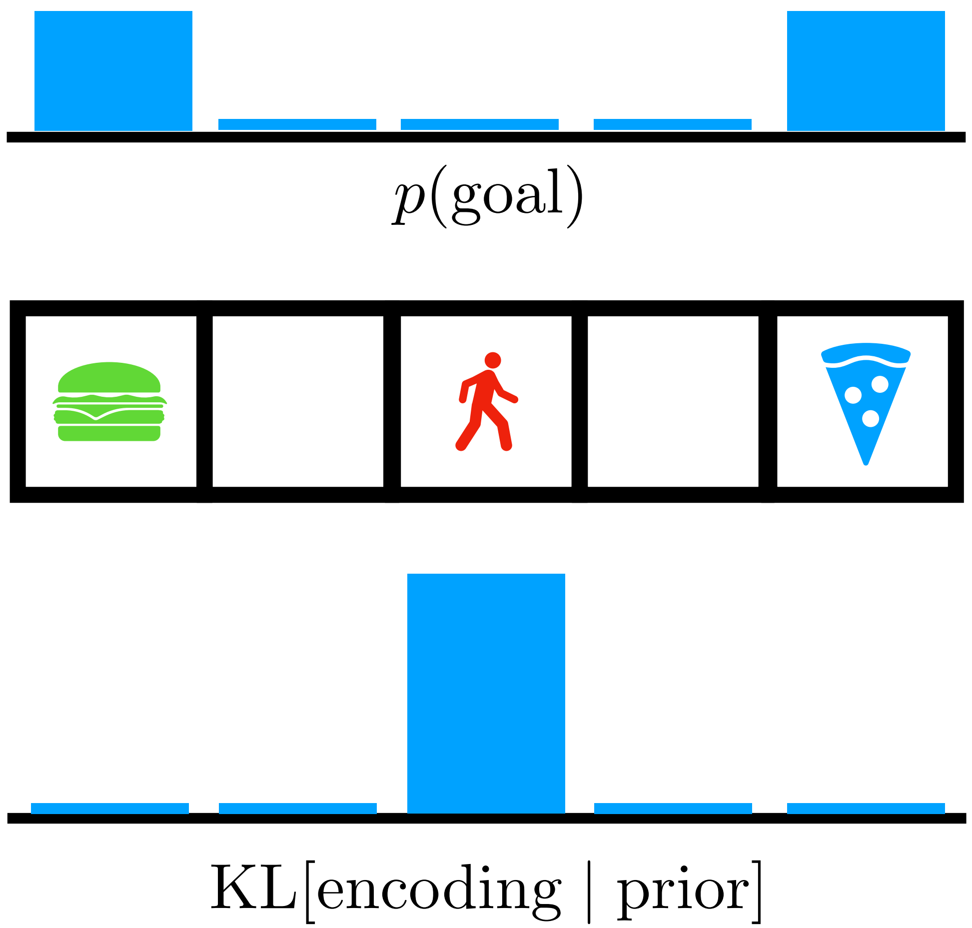
$$I(A; G \mid S) \leq I(Z; G \mid S)$$
$$\leq \underbrace{\sum_g p(g)}_{\text{sample a goal}} \underbrace{\sum_s p(s \mid g)}_{\text{sample trajectory}} \underbrace{\mathrm{KL}[p_{\text{enc}}(z \mid s, g) \mid r(z)]}_{\text{penalize encoder for departures from prior}}$$

additional details

- train using REINFORCE with state-value baseline
- 2 regularizations: above + entropy
- tabular encoder / decoder
- discrete latents (marginalized out; not sampled)
- fixed uniform prior

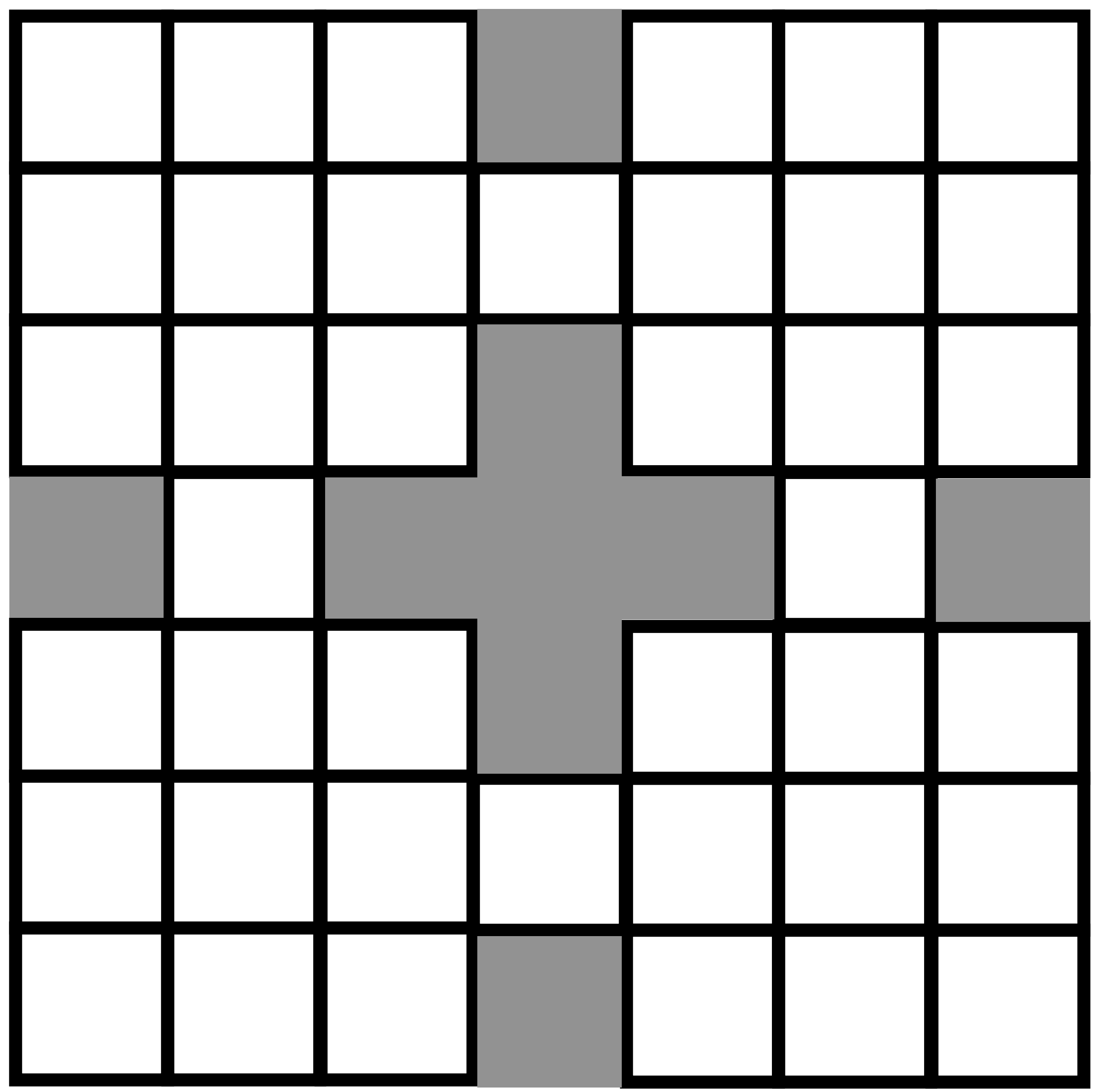
A simple example





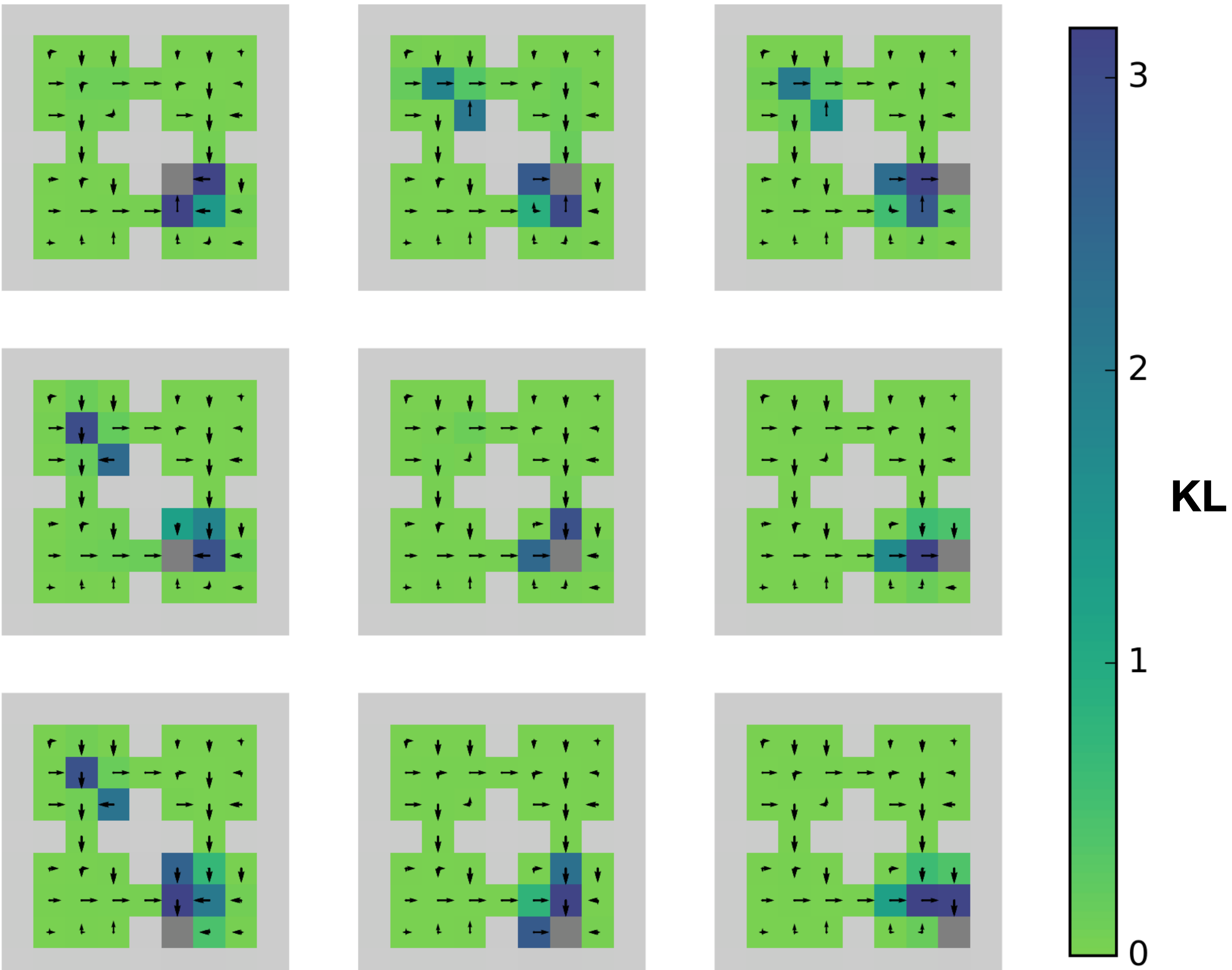
Four-room task

Agent spawns
in NW

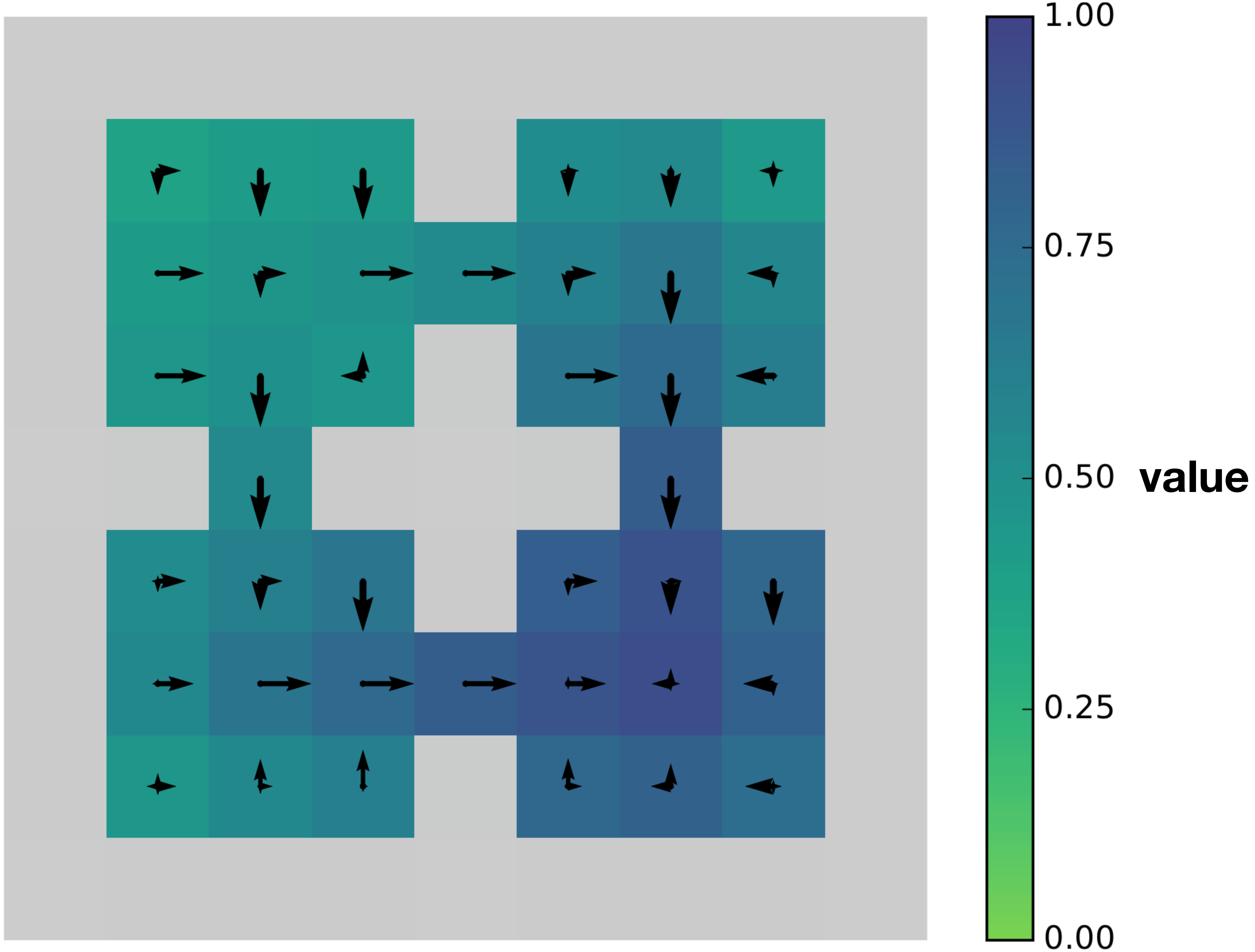


Goal spawns
in SE

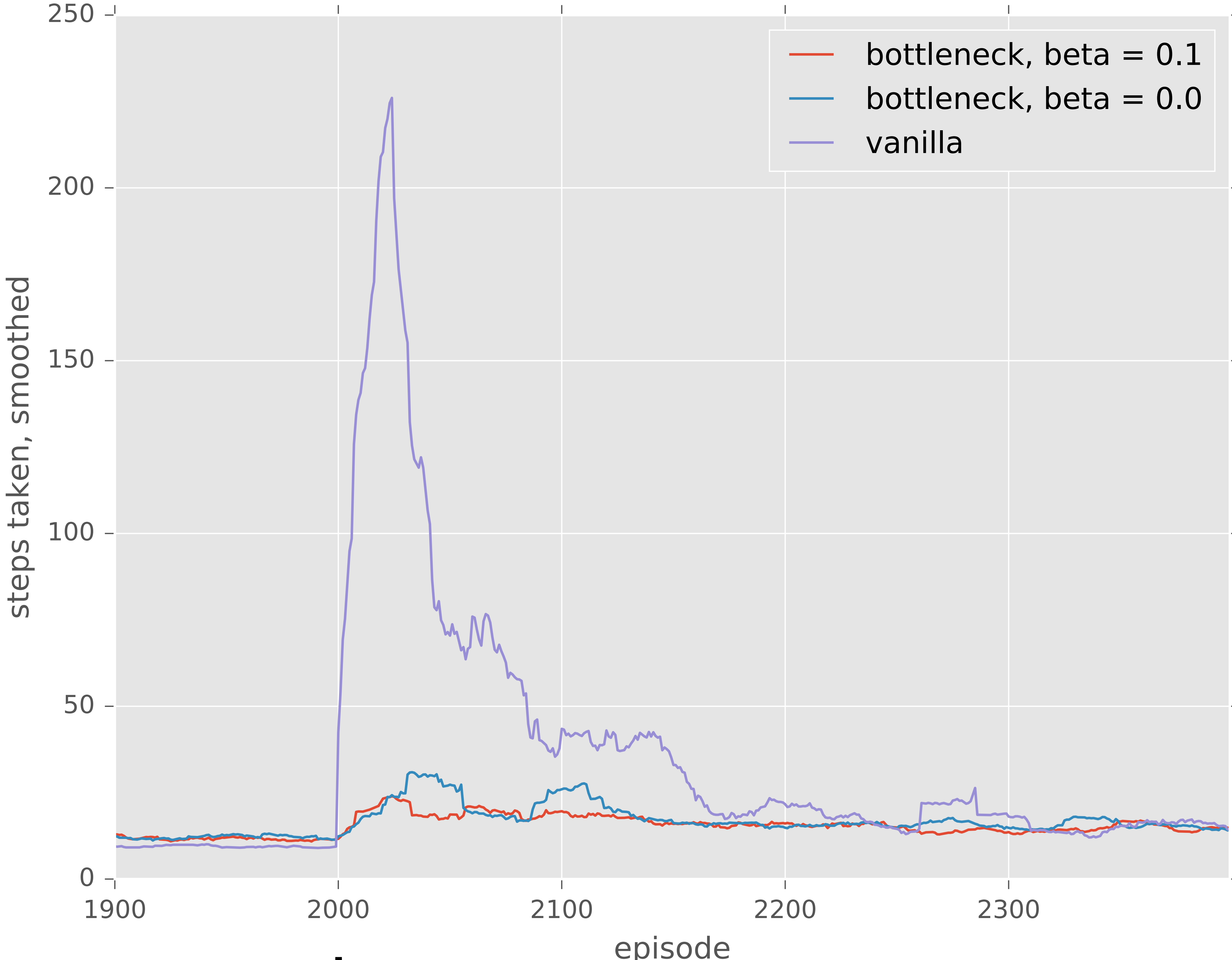
Results: agent learns selective goal lookup



Results: agent learns useful habits



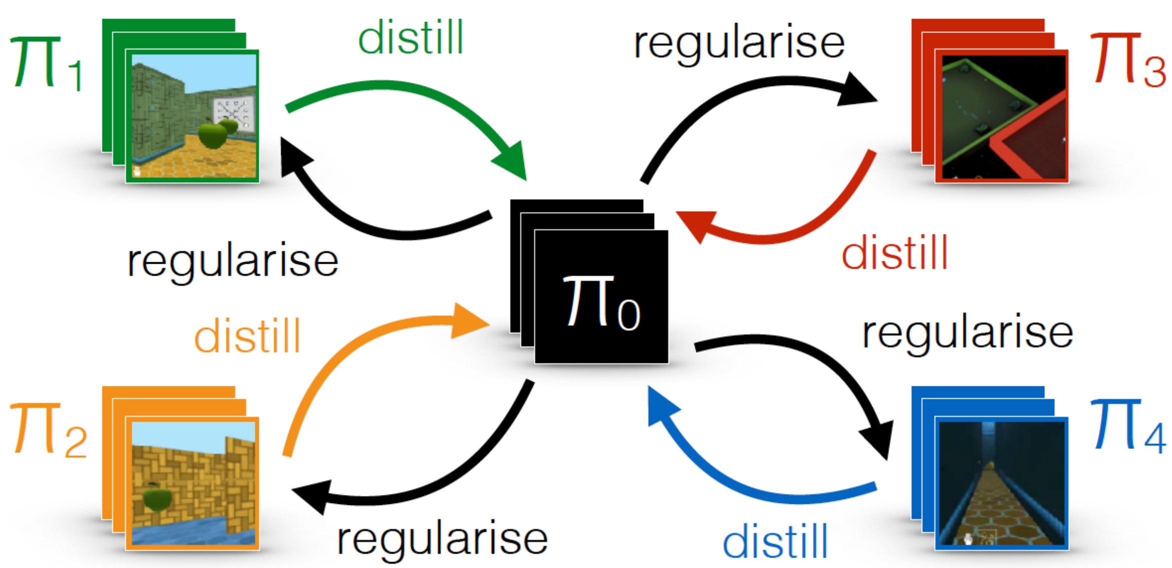
Results: agent transfers well



train on half the goals ← → **train on other half**

caveat: function approx would shrink this gap in transfer

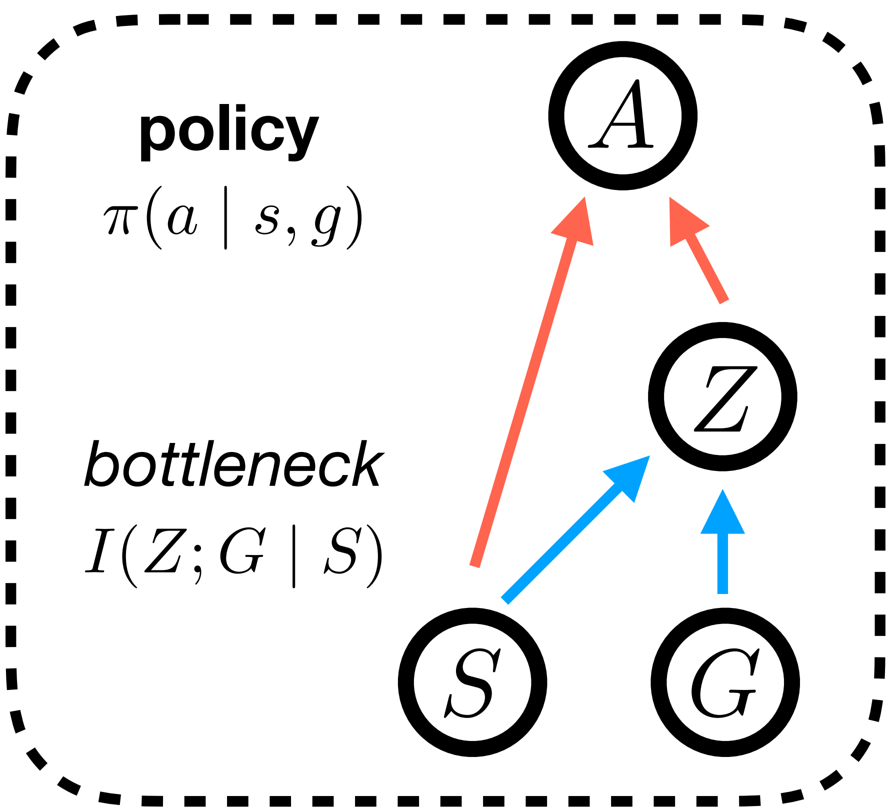
Results: equivalence to Distral



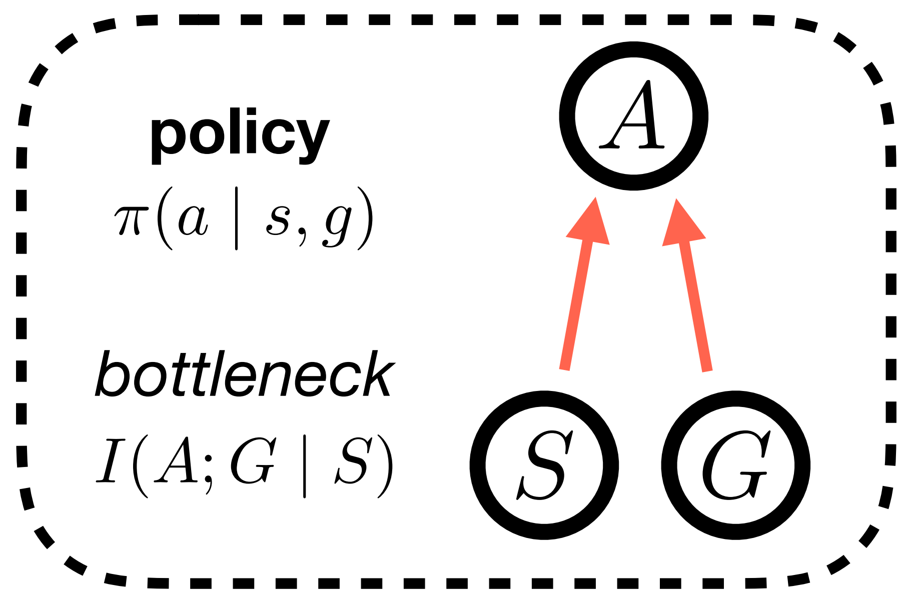
Distral: Robust Multitask Reinforcement Learning, Teh et al, NIPS 2017

main idea

regularize policy with $\mathrm{KL}[\pi_g \mid \pi_0]$



circumvent latents



$$I(A; G \mid S) \leq \sum_g p(g) \sum_s p(s \mid g) \mathrm{KL}[\pi_g(a \mid s) \mid \pi_0(a \mid s)]$$

Future directions

- **Slowly varying latents** (so that latents are endowed with meaning beyond single actions, i.e. trajectories)
- **Mixture models for base policy** (not entirely straightforward - need extra encouragement for the components to be meaningful)
- **Predictions for neuroscience / cognitive science** (e.g. agents should encode goal info only when needed, and distinguish between goals only to the extent it informs actions)
- **Alternative approaches to policy compression** (e.g. McNamee, Wolpert, & Lengyel 2016)
- **Using KL[encoding|prior] to prioritize experience replay** (or as target states for exploration under new goal)