Decoupling Exploration and Exploitation for Meta-Reinforcement Learning without Sacrifices

Evan Zheran Liu

Aditi Raghunathan

Percy Liang

Chelsea Finn

The world constantly **changes**

New kitchens / recipes

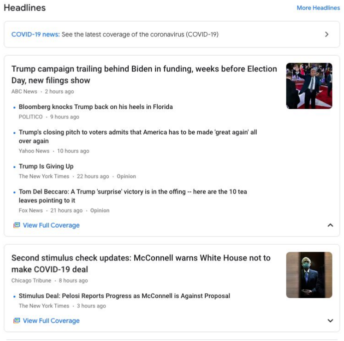
```
add $1, %rdi
mov %edx, %eax
shr $8, %rdx
xor -1(%rdi), %al
movzx %al, %eax
xor 0x4110a(, %rax, 8), %rdx
cmp %rcx, %rdi
```

Code

model

Prefetch `0x4110a`

New programs


Headlines
More Headlines

COVID-19 news: See the latest coverage of the coronavirus (COVID-19)

Trump campaign trailing behind Biden in funding, weeks before Election Day, new filings show
ABC News · 2 hours ago

- Bloomberg knocks Trump back on his heels in Florida
  POLITICO · 9 hours ago
- Trump's closing pitch to voters admits that America has to be made 'great again' all over again
  Yahoo News · 10 hours ago
- Trump Is Giving Up
  The New York Times · 22 hours ago · Opinion
- Tom Del Beccaro: A Trump 'surprise' victory is in the offing -- here are the 10 tea leaves pointing to it
  Fox News · 21 hours ago · Opinion

View Full Coverage

Second stimulus check updates: McConnell warns White House not to make COVID-19 deal
Chicago Tribune · 8 hours ago

- Stimulus Deal: Pelosi Reports Progress as McConnell is Against Proposal
  The New York Times · 3 hours ago

View Full Coverage

New users / preferences

The world constantly **changes**

New kitchens / recipes

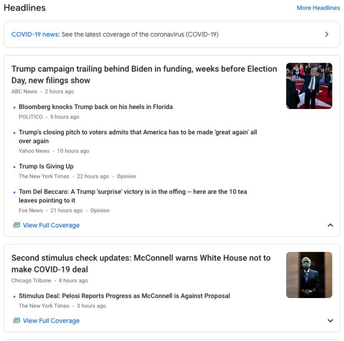
```
add $1, %rdi
mov %edx, %eax
shr $8, %rdx
xor -1(%rdi), %al
movzx %al, %eax
xor 0x4110a(, %rax, 8), %rdx
cmp %rcx, %rdi
```

Code

model

Prefetch `0x4110a`

New programs



New users / preferences

Re-learning from scratch is **expensive** and **sample inefficient**

Goal: quickly learn new tasks

How? Leverage *prior experience*

Goal: quickly learn new tasks

meta-training

How? Leverage *prior experience*

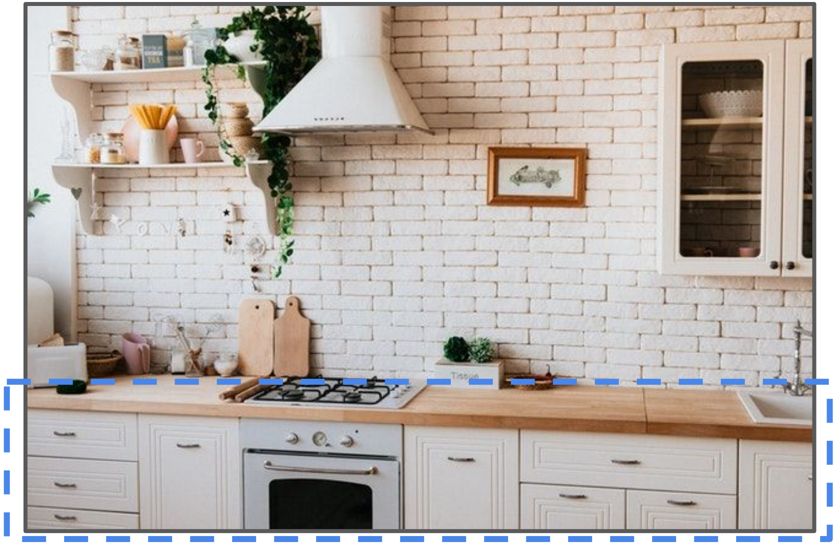
Goal: quickly learn new tasks

meta-training

meta-testing

How? Leverage *prior experience*

Goal: quickly learn new tasks



meta-training

explore

How? Leverage *prior experience*

Icons made by Freepik and ThoseIcons on flaticon.com

Goal: quickly learn new tasks

meta-training

exploit

How? Leverage *prior experience*

Icons made by Freepik and ThoseIcons on flaticon.com

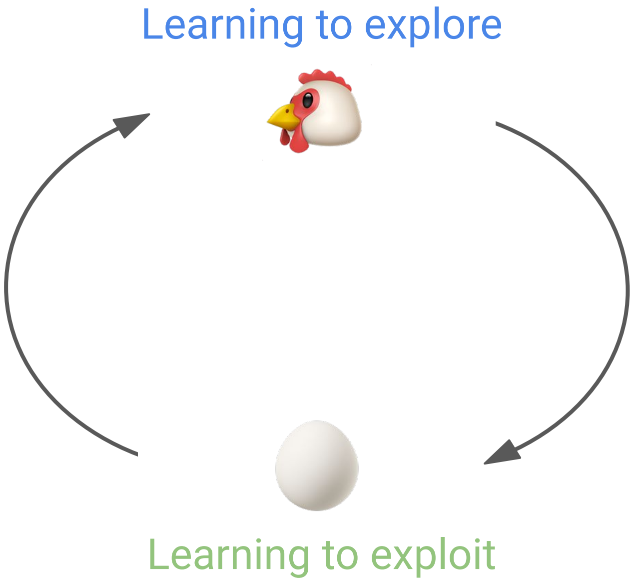
End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns

Duan et al., 2016; Wang et al., 2016; Mishra et al., 2017; Stadie et al., 2018; Zintgraf et al., 2019; Kamienny et al., 2020

End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns



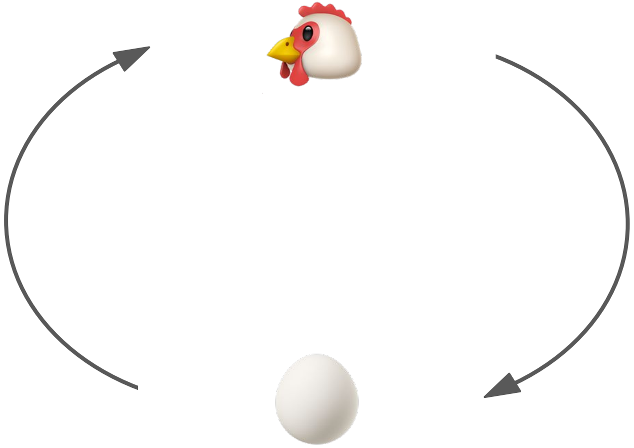
Coupling problem: learning exploration and exploitation depend on each other

Duan et al., 2016; Wang et al., 2016; Mishra et al., 2017; Stadie et al., 2018; Zintgraf et al., 2019; Kamienny et al., 2020

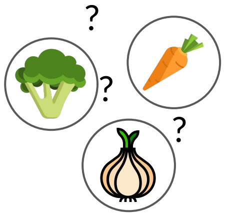
End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns

Learning to explore



Learning to exploit

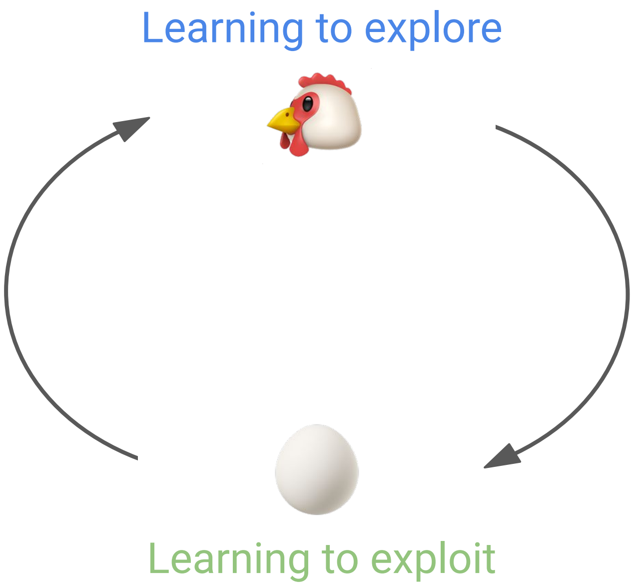


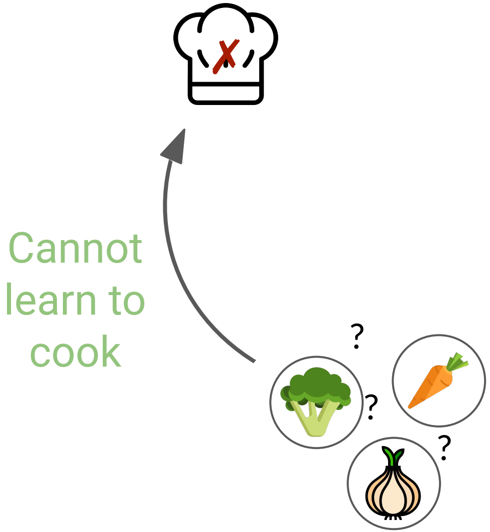
Coupling problem: learning exploration and exploitation depend on each other

Duan et al., 2016; Wang et al., 2016; Mishra et al., 2017; Stadie et al., 2018; Zintgraf et al., 2019; Kamienny et al., 2020

End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns



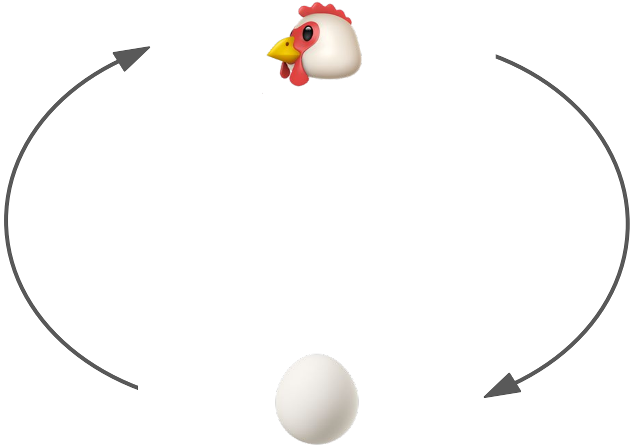


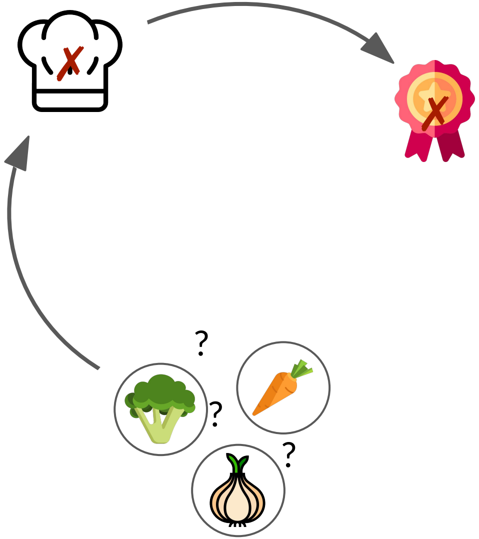
Coupling problem: learning exploration and exploitation depend on each other

Duan et al., 2016; Wang et al., 2016; Mishra et al., 2017; Stadie et al., 2018; Zintgraf et al., 2019; Kamienny et al., 2020

End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns



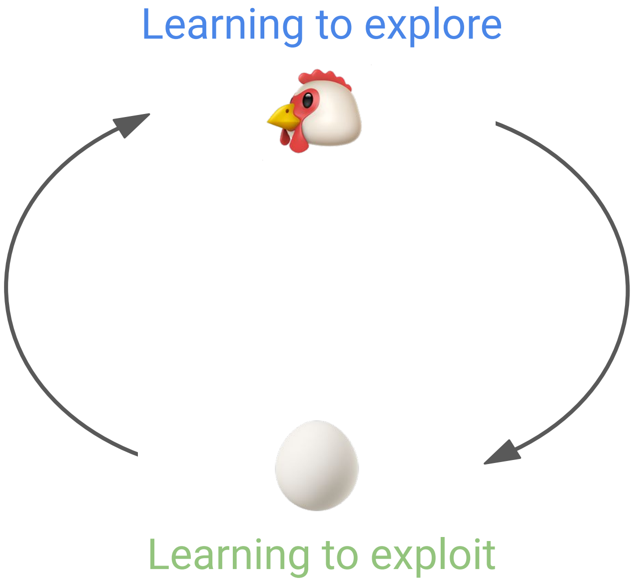


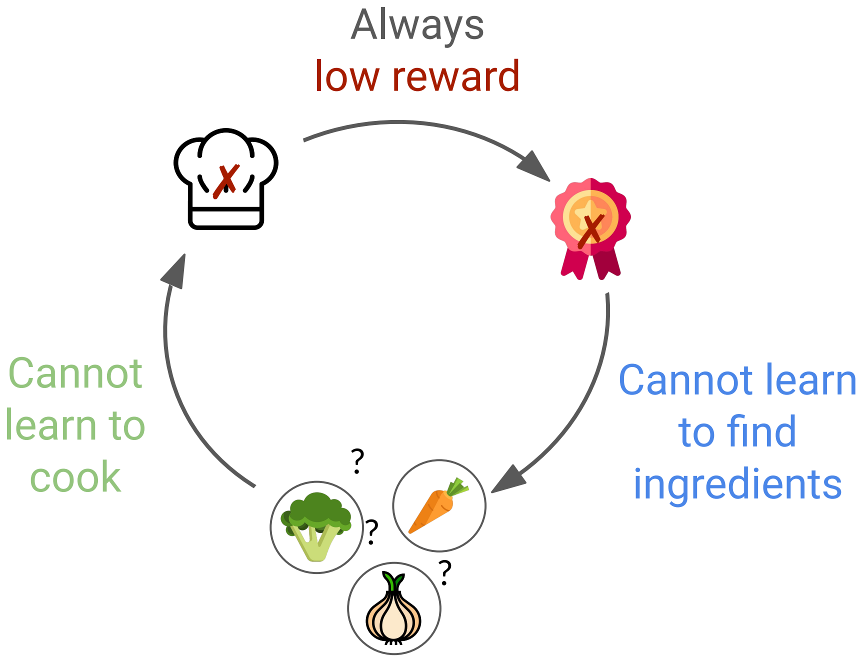
Coupling problem: learning exploration and exploitation depend on each other

Duan et al., 2016; Wang et al., 2016; Mishra et al., 2017; Stadie et al., 2018; Zintgraf et al., 2019; Kamienny et al., 2020

End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns



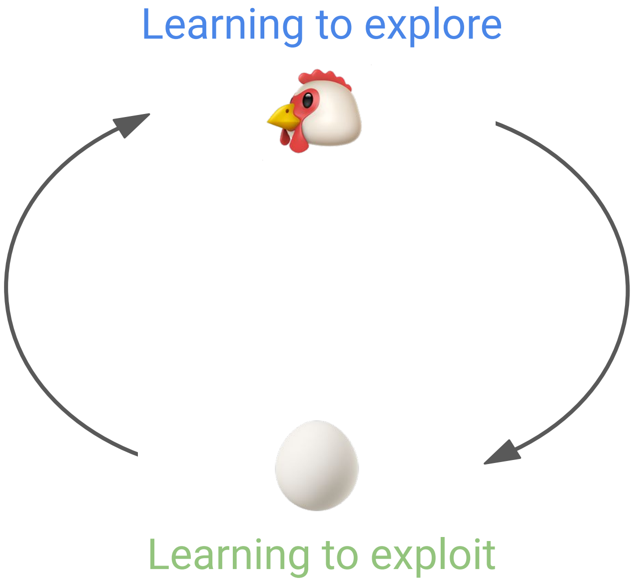


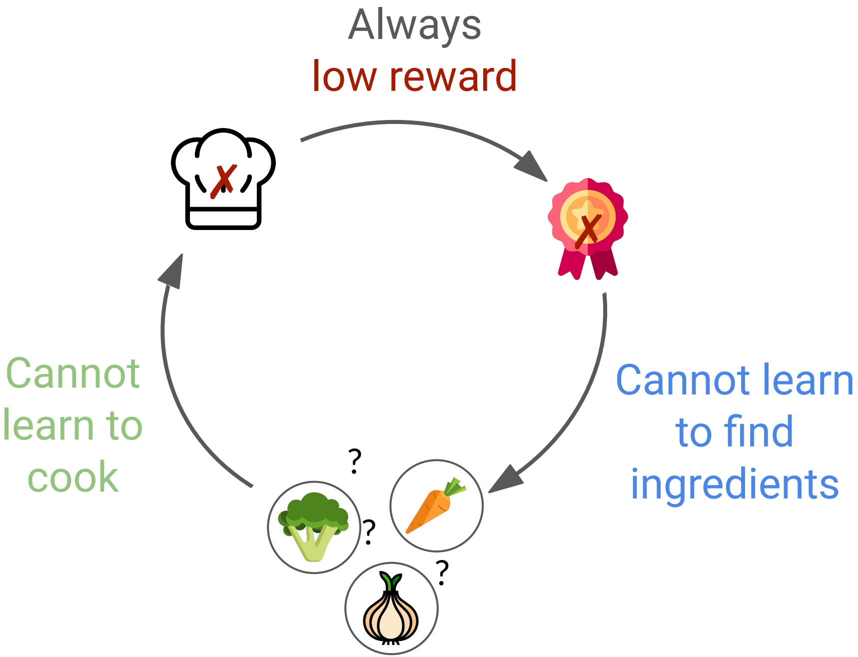
Coupling problem: learning exploration and exploitation depend on each other

Duan et al., 2016; Wang et al., 2016; Mishra et al., 2017; Stadie et al., 2018; Zintgraf et al., 2019; Kamienny et al., 2020

End-to-End Meta-Reinforcement Learning

Natural approach: Optimize exploration and exploitation **end-to-end** to maximize returns





Coupling problem: learning exploration and exploitation depend on each other

… leads to **local optima** and **poor sample complexity**

DREAM Overview

Goal: Create exploration objective to *all* and *only* recover **task-relevant** information

DREAM Overview

Goal: Create exploration objective to *all* and *only* recover **task-relevant** information

Key (mild) assumption: can distinguish all meta-training tasks from each other with a unique problem ID

DREAM Overview

Goal: Create exploration objective to *all* and *only* recover **task-relevant** information

Meta-training

1) Learn exploitation & identify task-relevant information

DREAM Overview

Goal: Create exploration objective to *all* and *only* recover **task-relevant** information

Meta-training

1) Learn exploitation & identify task-relevant information

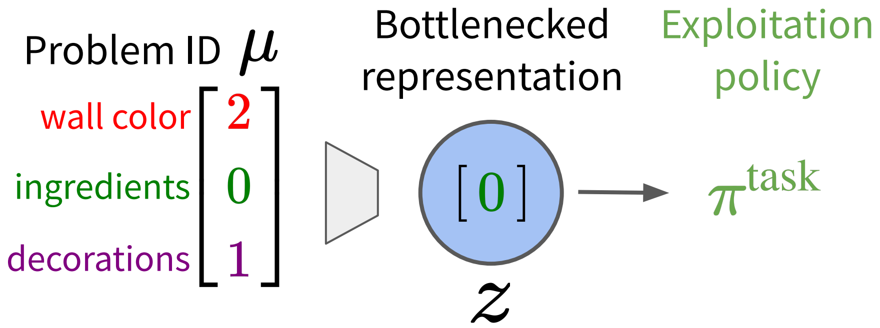
2) Learn to explore by recovering that information

DREAM Overview

Goal: Create exploration objective to *all* and *only* recover **task-relevant** information

Meta-training

1) Learn exploitation & identify task-relevant information



2) Learn to explore by recovering that information

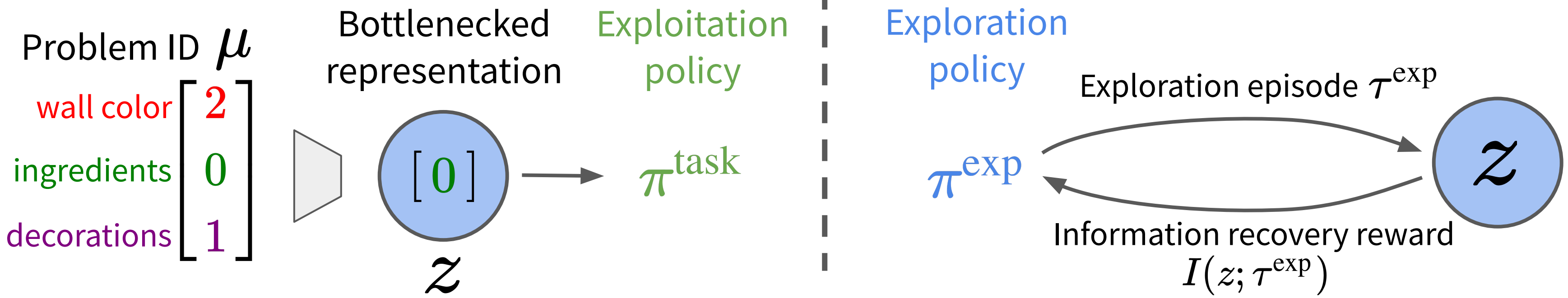
DREAM Overview

Goal: Create exploration objective to *all* and *only* recover **task-relevant** information

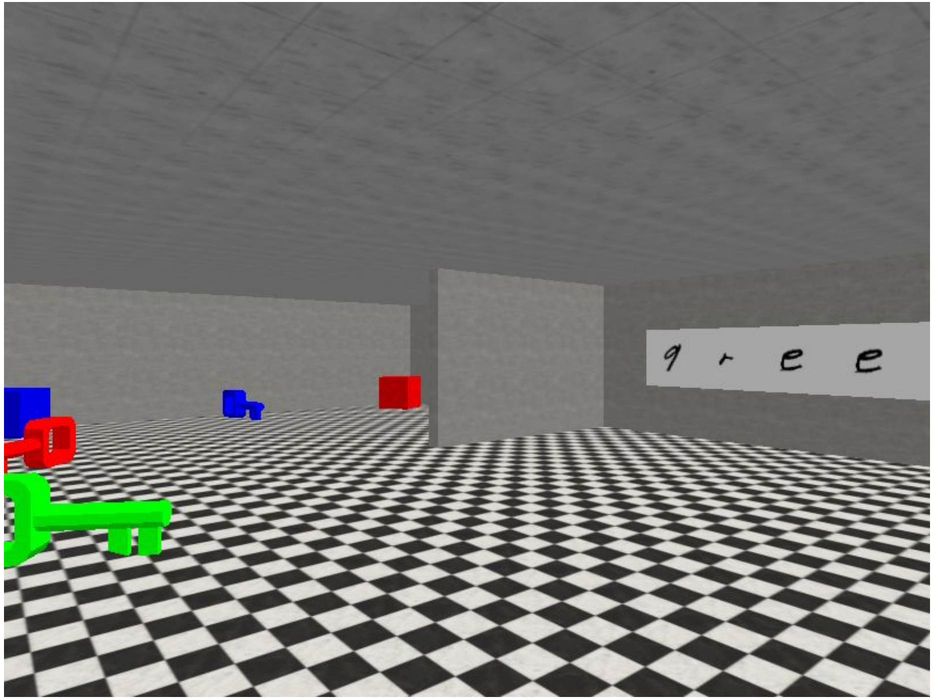
Meta-training

1) Learn exploitation & identify task-relevant information

2) Learn to explore by recovering that information



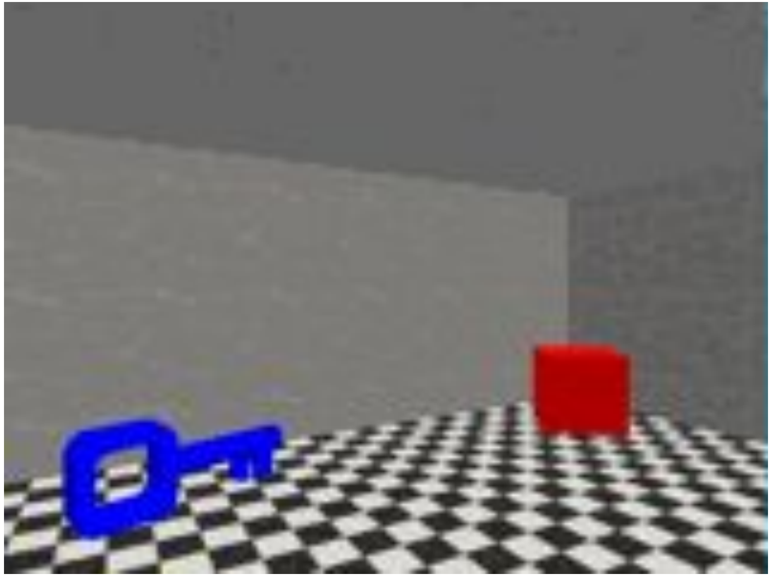
Experiments: Sparse Reward 3D Visual Navigation



More challenging variant of task from Kamienny et al., 2020

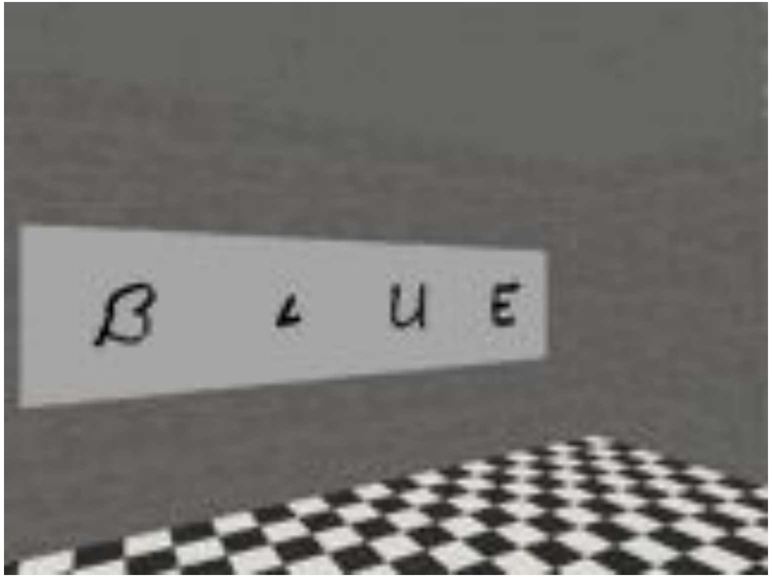
- Task: go to the *goal = key / block*, color specified by the sign
- Agent starts on other side of barrier and must walk around to read the sign
- Pixels observations (80 x 60 RGB)
- Sparse binary reward
- Existing benchmarks don't typically use **pixel observations** and **sparse rewards**

Experiments: Qualitative Results for DREAM

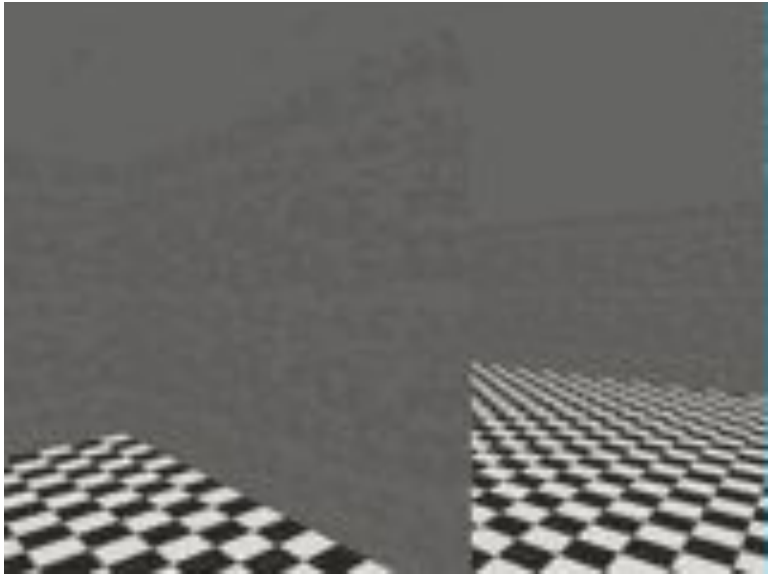


DREAM exploration

Experiments: Qualitative Results for DREAM

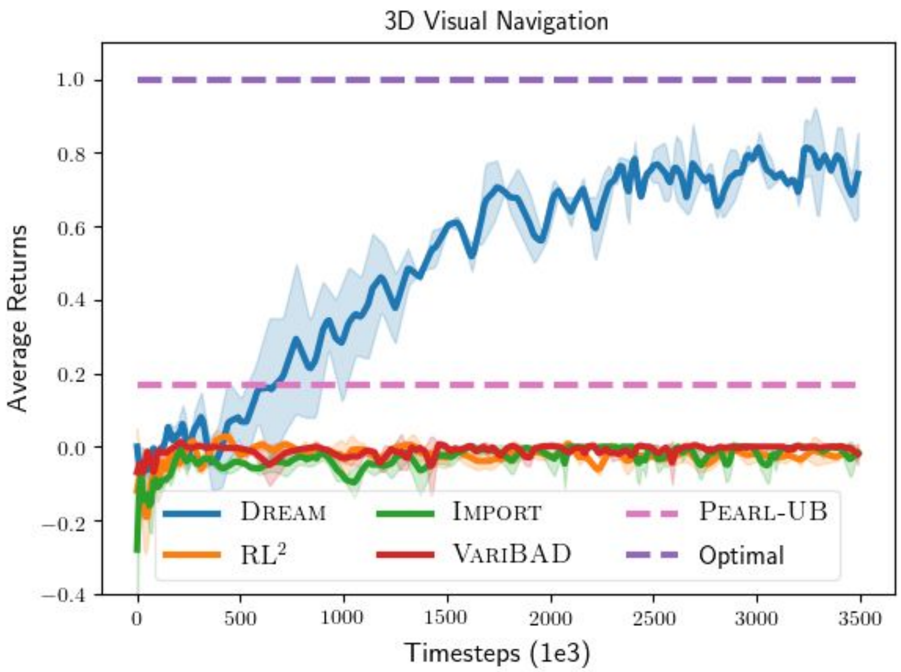


DREAM exploration



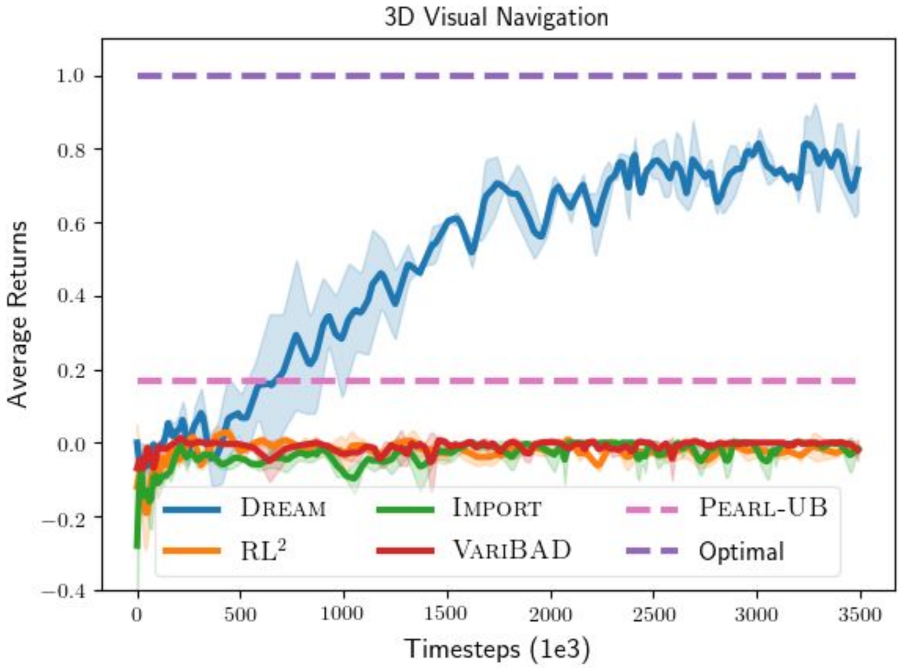
DREAM exploitation: Go to key

Experiments: Quantitative Results



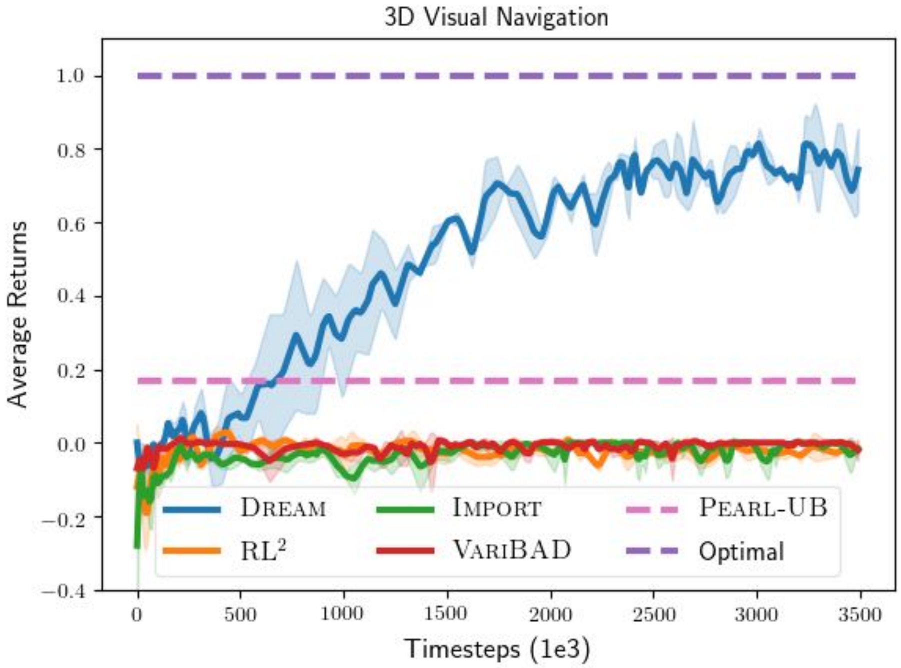
- Only DREAM scales to high-dimensional states and sparse rewards

Experiments: Quantitative Results



- Only DREAM scales to high-dimensional states and sparse rewards
- End-to-end approaches achieve *zero* returns due to **coupling problem**

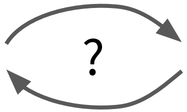
Experiments: Quantitative Results



- Only DREAM scales to high-dimensional states and sparse rewards
- End-to-end approaches achieve *zero* returns due to **coupling problem**
- Decoupled approaches, e.g., Thompson Sampling do not learn the optimal exploration strategy

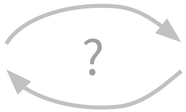
Takeaways

I. **Coupling** between exploration and exploitation prevents existing **end-to-end** methods from solving tasks with challenging exploration



Takeaways

I. **Coupling** between exploration and exploitation prevents existing **end-to-end** methods from solving tasks with challenging exploration



II. **DREAM** provides separate exploration and exploitation objectives that avoid the **coupling** problem