

<https://thegradient.pub/how-to-fix-rl/>

∇ The Gradient

- [HOME](#)
- [EDITOR'S NOTE](#)
- [OVERVIEWS](#)
- [PERSPECTIVES](#)
- [ABOUT](#)
- [SUBSCRIBE](#)
-

How to fix reinforcement learning

T

his piece is the second in a two-part series, starting with Reinforcement learning's foundational flaw.

In [part 1](#), we have already set up our board game allegory and demonstrated that pure RL techniques are limited

. In this part, we will enumerate various methods of incorporating prior knowledge and instruction into deep learning, and survey some amazing recent work into doing just that to conclude it is most definitely possible.

Why have we not moved beyond pure RL?

You might be thinking something like this:

We cannot just move beyond pure RL to emulate human learning — pure RL is rigorously formulated, and our algorithms for training AI agents are proven based on that formulation. Though it might be nice to have a formulation that aligns more closely with how people learn instead of learning-from-scratch, we just don't have one.

It's true that algorithms that incorporate prior knowledge or instructions are by definition more complex than the pure RL ones that have been rigorously formulated over decades. But that last bit is *not* true --- we do in fact have a formulation for learning-not-from-scratch which aligns more closely with how people learn.

Let's start by more explicitly describing how human learning is different from pure RL. When starting to learn a new skill, we basically do one of two things: guess at what the instructions might be (recall our prior experience with board games), or read some instructions (check the board game's rules). We generally know the goal and broad approach for a particular skill from the get-go, and we never reverse-engineer these things from a low-level reward signal.

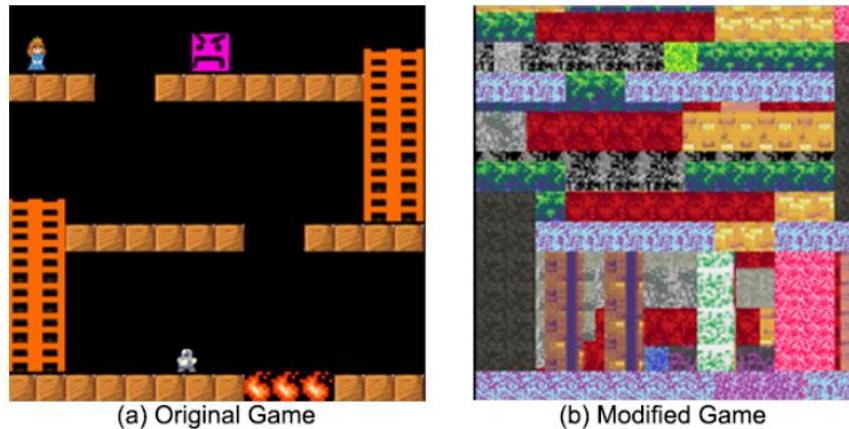


Figure 1: Motivating example. (a) A simple platformer game. (b) The same game modified by re-rendering the textures. Despite the two games being structurally the same, human players took twice as long to finish the second game as the first one. In comparison, the performance of an RL agent was approximately the same for the two games.

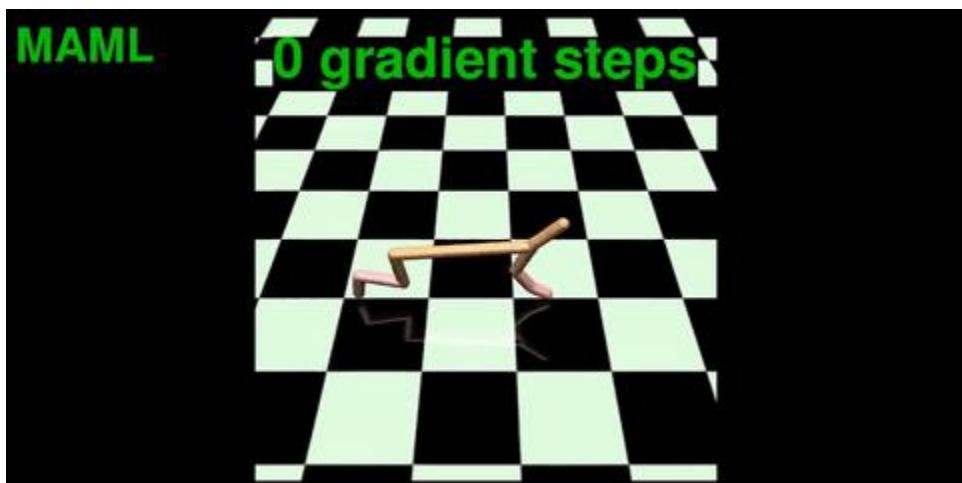
Researchers at UC Berkeley have recently demonstrated that humans learn much faster than pure RL in part due to making use of prior experience.

From Investigating Human Priors for Playing Video Games

Leveraging prior experience and instruction

The ideas of leveraging prior experience and getting instructions have very direct analogues in AI research:

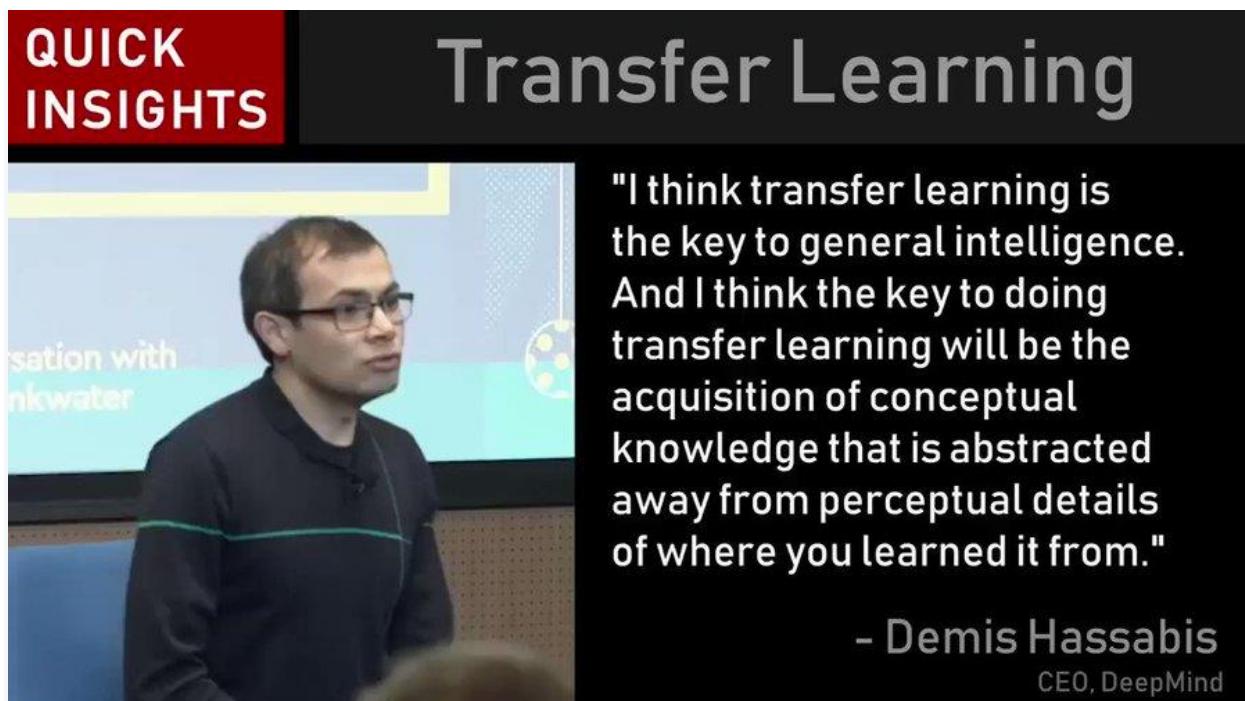
- **Meta-learning** tackles the problem of **learning how to learn**: making RL agents pick up new skills faster having already learned similar skills. And **learning how to learn**, as we'll see, is just what we need to move beyond pure RL and leverage prior experience.



A cutting-edge meta-learning algorithm, MAML. The agent is able to learn both backward

and forward running with very few iterations by leveraging meta-learning.
From "Learning to Learn"

- **Transfer learning**, roughly speaking, corresponds to 'transferring' skills attained in one problem to another potentially different problem. Here's Demis Hassabis, CEO of DeepMind, talking about the importance of transfer learning:



**QUICK
INSIGHTS**

Transfer Learning

"I think transfer learning is the key to general intelligence. And I think the key to doing transfer learning will be the acquisition of conceptual knowledge that is abstracted away from perceptual details of where you learned it from."

- Demis Hassabis
CEO, DeepMind



[Lex Fridman](#) @lexfridman

"I think transfer learning is the key to general intelligence. And I think the key to doing transfer learning will be the acquisition of conceptual knowledge that is abstracted away from perceptual details of where you learned it from." - Demis Hassabis
@demishassabis

270
[9:39 AM - Mar 17, 2018](#)

104 people are talking about this

And I think that [Transfer Learning] is the key to actually general intelligence, and that's the thing we as humans do amazingly well. For example, I played so many board games now, if someone were to teach me a new board game I would not be coming to that fresh anymore, **straight away I could apply all these different heuristics that I learned from all these other games to this new one even if I've never seen this one before, and currently machines cannot do that...** so I think that's actually one of the big challenges to be tackled towards general AI.

- **Zero-shot learning** is similar. It also aims to learn new skills fast, but takes it further by not leveraging **any** attempts at the new skill; the learning agent just receives 'instructions' for the new task, and is supposed to be able to perform well without any experience of the new task.
- **One-shot** and **few-shot** learning are also active areas of research. These fields differ from zero-shot learning in that they use demonstrations of the skill to be learned, or just a few iterations of experience, rather than indirect 'instructions' that do not involve the skill actually being executed.
- **Life Long Learning** and **Self Supervised Learning** are yet more examples of learning, in roughly defined as long-term continuous learning without human guidance.

	TRANSFER LEARNING	MULTI TASK LEARNING	FEW SHOT LEARNING	META LEARNING
LEARNING WITH LESS DATA	✓	✓	✓	✓
SOLVING DIVERSE PROBLEMS	-	✓	-	✓
ADJUSTING THE INCREMENTAL DATA	-	-	✓	✓

From [Effective Learning: The Near Future of AI](#)

These are all methodologies that go beyond learning from scratch. In particular, **meta-learning and zero-shot learning capture different elements of how a human would actually approach that new board**

game situation. A meta-learning agent would leverage experience with prior board games to learn faster, though it would not ask for the rules of the game. On the other hand, a zero-shot learning agent would ask for the instructions, but then not try to do any learning to get better beyond its initial guess of how to play the game well. One- and few-shot learning incorporate parts of both, but are limited by only getting demonstrations of how the skill can be done — that is, the agent would observe others playing the board game, but not request explanations or the rules of the game.

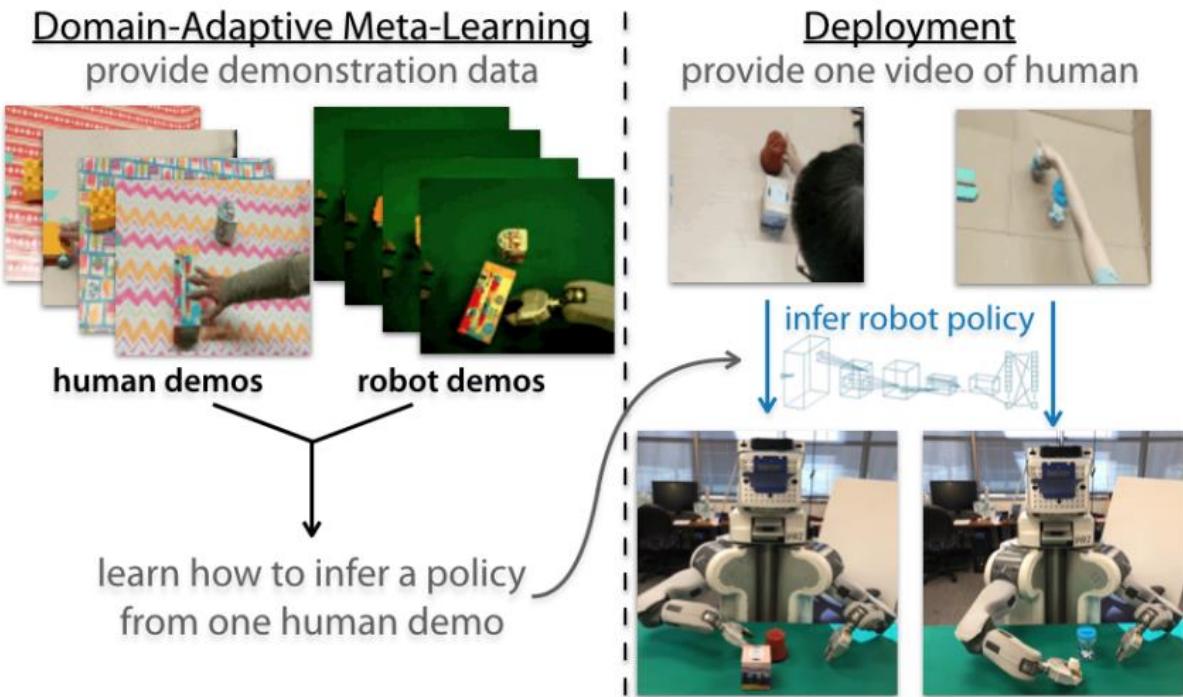


Fig. 1. After meta-learning with human and robot demonstration data, the robot learns to recognize and push a new object from one video of a human.

A recent 'hybrid' approach that combines one-shot and meta-learning.

From "[One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning](#)".

The broad notions of meta-learning and zero/few-shot learning are what 'make sense' in the context of the board game allegory. Better yet, hybrids of zero/few-shot and meta learning come close to representing what people actually do. They use prior experience, instructions, and trial runs to form an initial hypothesis of how the skill should be done. Then, they actually try doing the skill themselves and rely on the reward signal to test and fine-tune their ability to do the task beyond this initial hypothesis.

It is therefore surprising that 'pure RL' approaches are still so predominant and research on meta-learning and zero-shot learning is less championed.

Part of the reason for this could be that the basic formulation of RL has not been questioned more, and that the notions of meta-learning and zero-shot learning have not been popularly encoded into its basic equations. Among research that *has* suggested alternative formulations of RL, perhaps the most relevant piece is DeepMind's 2015 "[Universal Value Function Approximators](#)", which generalized the idea of 'General value functions' introduced by Richard Sutton (by far the most influential researcher in RL) and collaborators in 2011. DeepMind's abstract summarizes the idea well:

"Value functions are a core component of [RL] systems. The main idea is to construct a single function approximator $V(s; \theta)$ that estimates the long-term reward from any state s , using parameters θ . In this paper we introduce universal value function approximators (UVFAs) $V(s, g; \theta)$ that generalise not just over states s but also over goals g ."

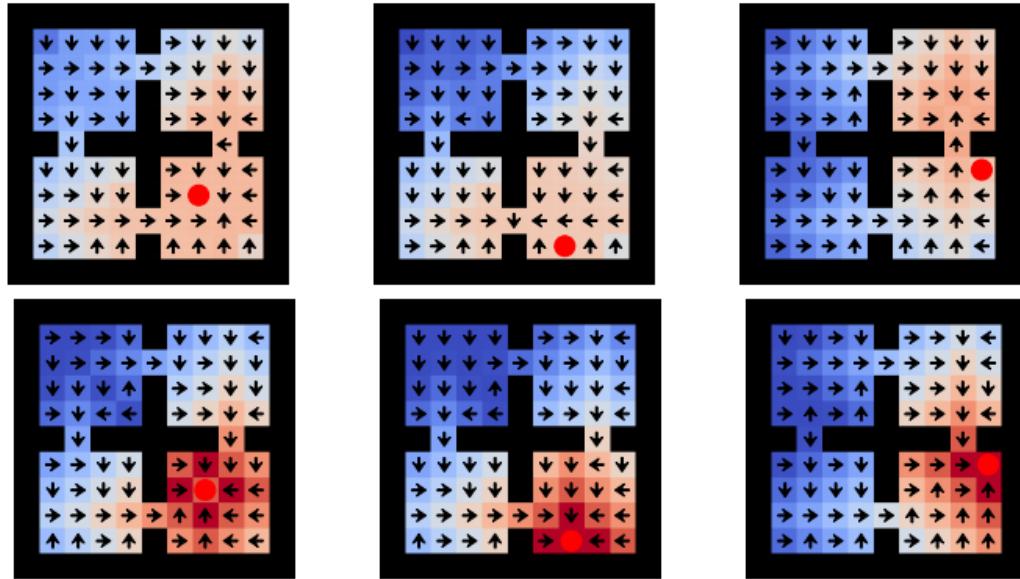


Figure 8. A UVFA trained only on goals in the first 3 rooms of the maze generalizes such that it can achieve goals in the fourth room as well. **Above:** This visualizes the learned UVFA for 3 test goals (marked as red dots) in the fourth room. The color encodes the value function, and the arrows indicate the actions of the greedy policy. **Below:** Ground truth for the same 3 test goals.

This UVFA idea put to practice. From "[Universal Value Function Approximators](#)".

Here is a rigorous, mathematical formulation of RL that treats goals (the high-level objective of the skill to be learned, which should yield good rewards) as a fundamental and necessary input rather than something to be discovered from just the reward signal. The agent is told what it's supposed to do, just as is done in zero-shot learning and actual human learning.

It has been 3 years since this was published, and how many papers have cited it since? **72**. A tiny fraction of all papers published in RL; for context, DeepMind's "Human-level control through deep RL" was also published in 2015 and as of now has **2906** citations, and their 2016 "Mastering the game of Go with deep neural networks and tree search" has **2882** citations according to Google Scholar.

So, work is definitely being done towards this notion of incorporating meta-learning/zero-shot learning with RL. But as shown by these citation counts, this research direction is still relatively obscure. Here is the key question: why is RL that incorporates meta-learning and/or zero-shot learning, as formalized by DeepMind's work, not the default?

To some extent the answer is obvious: it's hard. AI research tends to tackle isolated, well-defined problems in order to make progress on them, and there is less work on learning that strays from pure RL and learning from scratch precisely because it is harder to define. But, this answer is not satisfactory: deep learning has enabled researchers to create hybrid approaches, such as models that contain both Natural Language Processing and Computer Vision or for that matter the original AlphaGo's approach of combining both classic techniques and Deep-Learning for playing Go extremely well. In fact, DeepMind's own recent position paper "Relational inductive biases, deep learning, and graph networks" stated this point well:

"We suggest that a key path forward for modern AI is to commit to combinatorial generalization as a top priority, and we advocate for integrative approaches to realize this goal. Just as biology does not choose between nature versus nurture—it uses nature and nurture jointly, to build wholes which are greater than the sums of their parts—we, too, reject the notion that structure and flexibility are somehow at odds or incompatible, and embrace both with the aim of reaping their complementary strengths. In the spirit of numerous recent examples of principled hybrids of structure-based methods and deep learning (e.g., Reed and De Freitas, 2016; Garnelo et al., 2016; Ritchie et al., 2016; Wu et al., 2017; Denil et al., 2017; Hudson and

Manning, 2018), we see great promise in synthesizing new techniques by drawing on the full AI toolkit and marrying the best approaches from today with those which were essential during times when data and computation were at a premium."

Recent work on meta-learning / zero-shot learning

We now state our conclusion:

Motivated by the board game allegory, we should reconsider the basic formulation of RL along the lines of DeepMind's Universal Value Function idea, or at least double down on the already ongoing research that is implicitly doing just that through meta-learning, zero-shot learning, and more.

Much, if not most, of modern RL research still builds on pure RL approaches that leverage only the reward signal and possibly a model. Not only that, but the majority of attention is still given to such work, with the previously discussed [AlphaGo Zero](#) receiving more attention and praise than most recent AI work. The paper in which it was introduced "[Mastering the game of Go without human knowledge](#)", was published just last year and already has **406** citations; DeepMind's Universal Value Function paper has been published for thrice longer and has about a third of the citations with just **72**. Other notable papers that combine meta-learning and reinforcement learning stand at similar numbers: "[Learning to reinforcement learn](#)" has 58 citations and "[RL2: Fast Reinforcement Learning via Slow Reinforcement Learning](#)" has 52. But among those citations is some **very** exciting work:

- "[Hindsight Experience Replay](#)"
- "[Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning](#)"

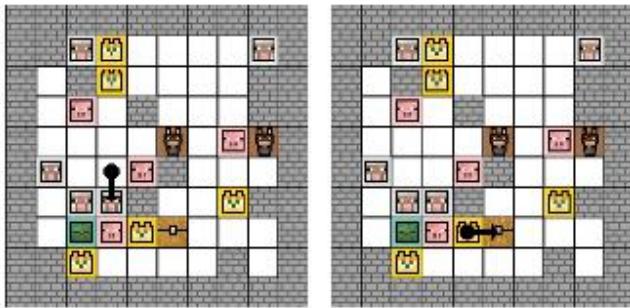
First-person-view
(Observation)



Training

1. Visit pig
2. Pick up 3 sheep
3. Transform greenbot
4. Pick up horse

Top-down-view
(Not available)

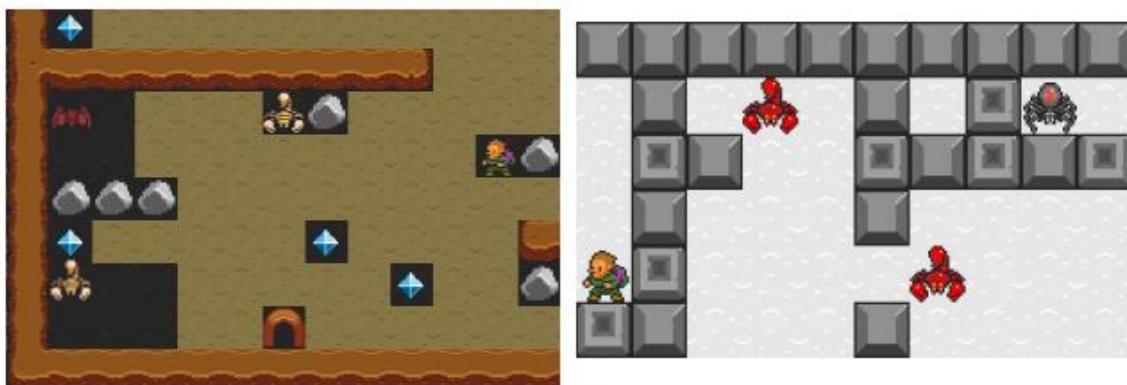


Testing

1. Pick up pig
 2. Visit sheep
 3. Transform cat
 4. Transform 3 sheep
 5. Pick up greenbot
 6. Pick up 2 pig
 7. Transform 2 sheep
 8. Transform 2 cat
- ...

From "[Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning](#)".

- ["Representation Learning for Grounded Spatial Reasoning"](#)
- ["Deep Transfer in Reinforcement Learning by Language Grounding"](#)



is an enemy who chases you



is a randomly moving enemy



is a stationary collectible



is a stationary immovable wall

Boulderchase

Bomberman

Figure 1: Two different game environments with a few associated text descriptions. Entity names are replaced with icons for the purpose of illustration.

From "[Deep Transfer in Reinforcement Learning by Language Grounding](#)".

- "[Cross-Domain Perceptual Reward Functions](#)"
- "[Learning Goal-Directed Behaviour](#)"

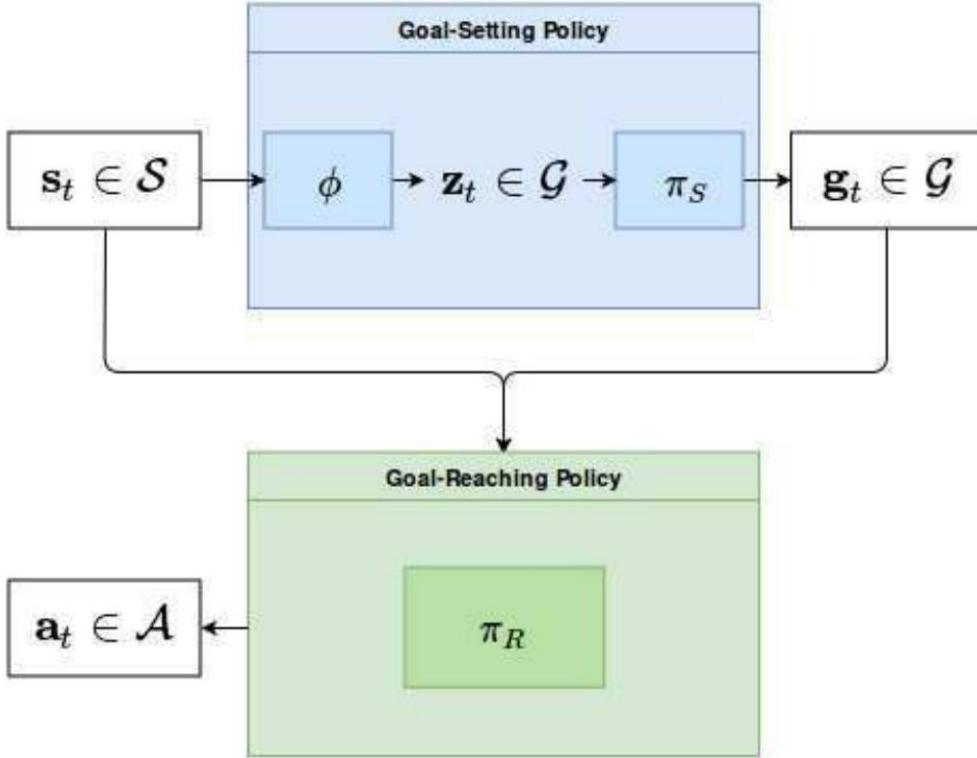


Figure 4.1: Illustration of the goal-based framework proposed in this thesis. Detailed information can be found in the corresponding subsections of this chapter.

From "Learning Goal-Directed Behaviour"

Now that's exciting! And, all these goal-specification/hybrid meta/zero/one shot approaches are arguably just the most obvious of directions to pursue for more human-inspired AI methods. Possibly even more exciting is the recent swell of work exploring intrinsic motivation and curiosity-driven exploration for learning (often motivated, curiously, by the way human babies learn):

- "Kickstarting Deep Reinforcement Learning"

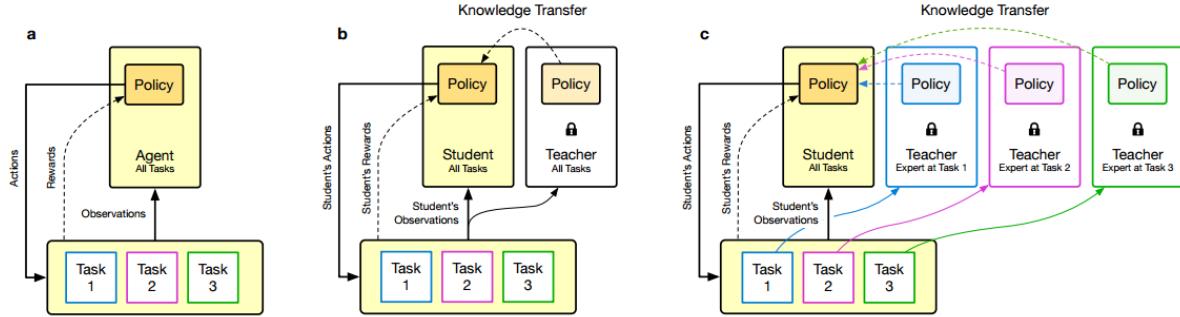


Figure 1. Overview. Schematics of the three agent training scenarios explored in this paper. **a:** Standard multi-task RL. An agent learns to perform well on several tasks by acting on and observing all of them. **b:** Single-teacher kickstarting (section 5.1). Observations from all tasks are also fed to a fixed, previously-trained teacher, and knowledge is transferred from teacher to student by encouraging the student to match the teacher’s actions. **c:** Multiple-teacher kickstarting (section 5.2). For each task, observations are sent to a task-specific expert teacher; the student is encouraged to match this teacher’s actions on the task.

From "Kickstarting Deep Reinforcement Learning"

- "Surprise-Based Intrinsic Motivation for Deep Reinforcement Learning"
- "Meta-Reinforcement Learning of Structured Exploration Strategies":

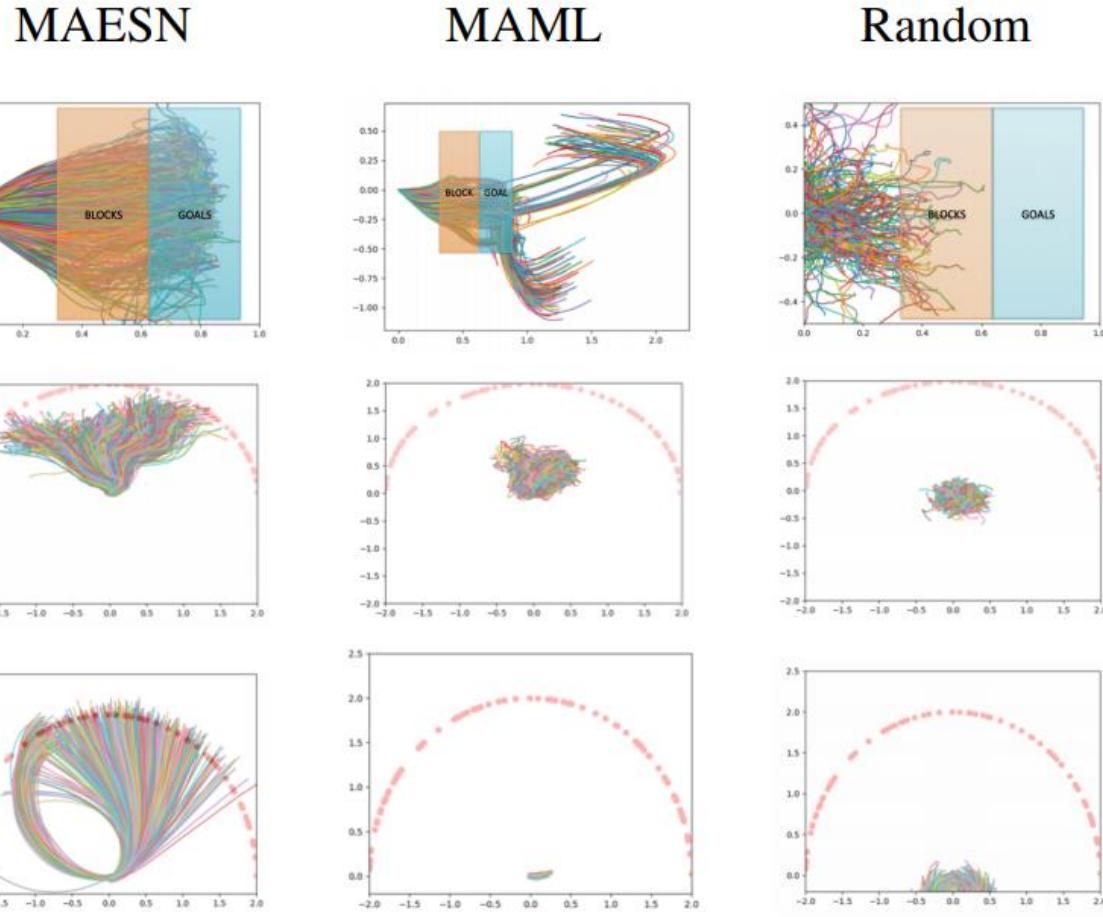


Figure 7. Plot of exploration behavior visualizing 2D position of the manipulator (for blockpushing) and the CoM for locomotion for MAESN, MAML and random initialization. **Top:** Block Manipulation **Middle:** Ant Locomotion **Bottom:** Wheeled Locomotion. Goals are indicated by the translucent overlays. We see that MAESN better captures the task distribution than other methods.

From "[Meta-Reinforcement Learning of Structured Exploration Strategies](#)"

- "[Learning Robust Rewards with Adversarial Inverse Reinforcement Learning](#)"
- "[Curiosity-driven Exploration by Self-supervised Prediction](#)"
- "[Learning by Playing - Solving Sparse Reward Tasks from Scratch](#)"
- "[Learning to Play with Intrinsically-Motivated Self-Aware Agents](#)"

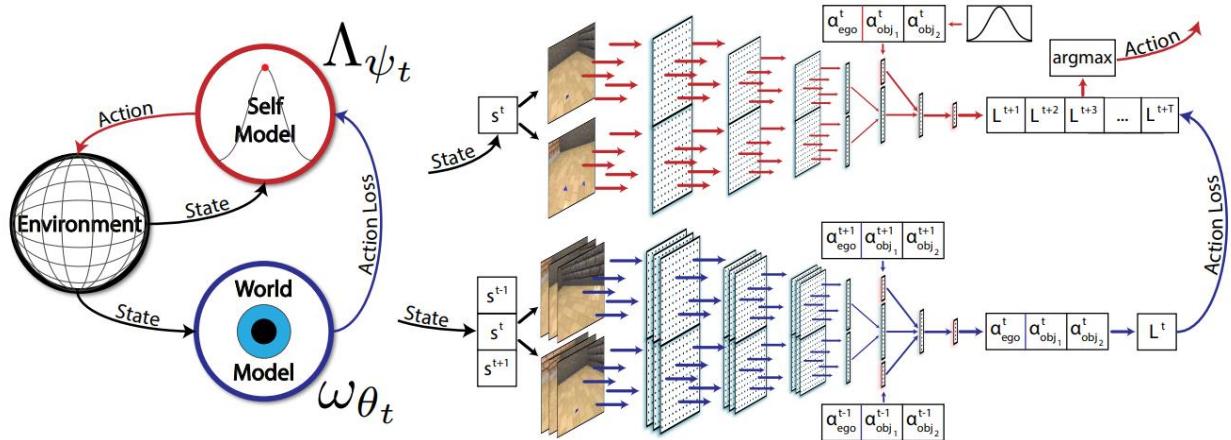


Figure 2. Intrinsically-motivated self-aware agent architecture. The world-model (blue) solves a dynamics prediction problem. Simultaneously a self-model (red) is learned that seeks to predict the world-model's loss. Actions are chosen to antagonize the world-model, leading to novel and surprising events in the environment (black).

From "Learning to Play with Intrinsically-Motivated Self-Aware Agents"

- "Unsupervised Predictive Memory in a Goal-Directed Agent"
- "World Models"

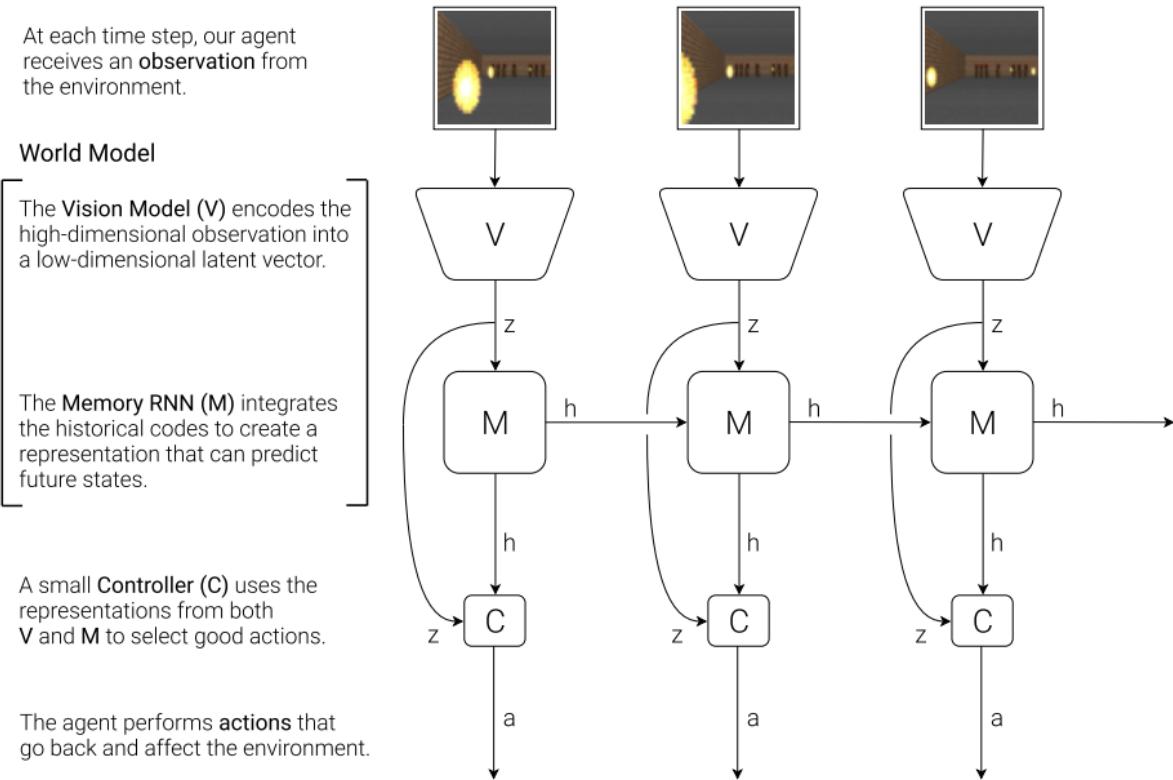


Figure 4. Our agent consists of three components that work closely together: **Vision (V)**, **Memory (M)**, and **Controller (C)**

From "World Models"

And we can even go beyond taking inspiration from human learning: we can directly study it. In fact, both older and cutting edge neuroscience research directly suggest human and animal learning can be modeled as reinforcement learning mixed with meta-learning:

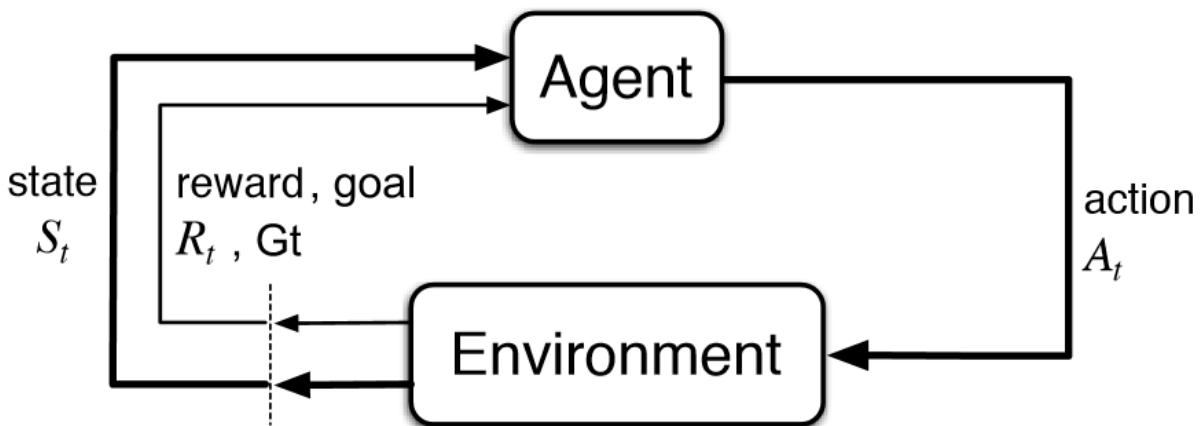
- "Meta-learning in Reinforcement Learning"
- "Prefrontal cortex as a meta-reinforcement learning system"
- :

The results of that last paper, “Prefrontal cortex as a meta-reinforcement learning system”, are particularly intriguing for our conclusion. Basically, one can even argue that human intelligence is powered at its very core by a combination of reinforcement learning and meta learning - **meta-reinforcement learning**. If that is the case, should we not do the same for AI?

Conclusion

The classic formulation of RL has a fundamental flaw that may limit it from solving any truly complex problems: its implied assumptions of starting from scratch and its heavy dependence on a low-level reward signal or provided environment model. As shown by the many papers cited here, going beyond starting from scratch does not necessitate hand-coded heuristics or rigid rules. Meta-RL methods empower AI agents to learn better through high-level instructions, accumulated experience, examples of what it should learn to do, learning a model of the world, intrinsic motivation, and more.

Let's end on an optimistic note: the time is ripe for the AI community to embrace work such as the above, and **move beyond pure RL** with more **human-inspired** learning approaches. Based on the **board-game allegory** alone, it seems reasonable to claim that AI techniques must move towards this and away from pure RL in the long term. Work on pure RL should not immediately stop, but it should be seen as useful *insofar as it is complementary to non-pure RL methods* and as long as we remain cognizant of its *inherent limitations*. If nothing else, methods based on meta-learning, zero/few-shot learning, transfer learning, and hybrids of all of these should become the default rather than the exception. And as a researcher about to embark on my PhD, I for one am willing to bet my most precious resource — my time — on it.



Is this really the future of RL?

Andrey Kurenkov is a graduate student affiliated with the Stanford Vision Lab, and lead editor of [Skynet Today](#). These opinions are solely his.