https://thegradient.pub/why-rl-is-flawed/

Reinforcement learning's foundational flaw

08.JUL.2018

In this essay, we are going to address the limitations of one of the core fields of AI.

In the process, we will encounter a fun allegory, a set of methods of incorporating prior knowledge and instruction into deep learning, and a radical conclusion.

The first part, which you're reading right now, will set up what RL is and why it (or at least a particular version of it we shall name 'pure RL' and soon define) is fundamentally flawed. It will contain some explanation that can be skipped by AI practitioners -- but be sure to stick around for the discussion of recent non pure-RL work we shall argue represents the fix to pure RL's foundational flaw. But for now, let us start with a fun allegory.

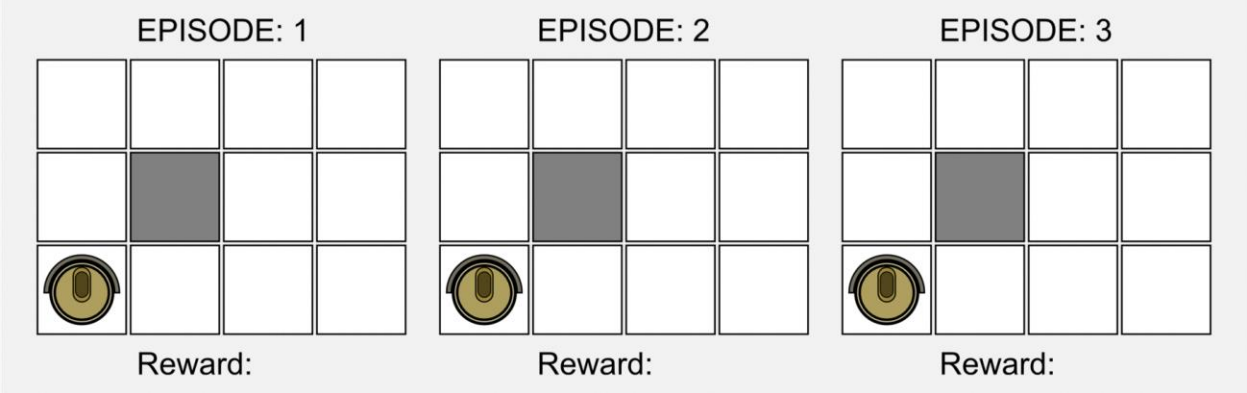
The allegory of the board game

Imagine this: a friend of yours invites you to play a board game you have never played. In fact, you have never played a board game, nor any kind of game, in your life. Your friend tells you what the valid moves are, but not what they mean or how the game is scored. So, you start playing — no more questions, no more explanations. And you lose. And lose. And lose. And… lose some more. Slowly you figure out some patterns in your losses; you're still losing, but not as quickly anymore. After a few weeks of consecutive play and many thousands of games, you even manage to just barely win.

Silly, right? Why didn't you just ask what the goal of the game is and how it is supposed to be played? Yet the above paragraph describes how the majority of reinforcement learning methods still work today.

Reinforcement learning (RL) is one of the basic subfields within AI. In an RL framework, an **agent** interacts with an **environment** to learn what **actions** it needs to take in any given environment **state** to maximize its long-term **reward**. In the board game allegory, this translates to having **you** interact with **the board** to learn what **moves** you should take in each **board game configuration** to maximize your **final score**.

In the typical model of RL, the agent begins only with knowledge of which actions are possible; it knows nothing else about the world, and it's expected to learn the skill solely by interacting with the environment and receiving rewards after every action it takes. The lack of prior knowledge means that the agent learns 'from scratch'. Let's call this learning-from-scratch approach **pure RL**. Pure RL has notably been used to tackle games like backgammon and Go, as well as various problems in robotics and elsewhere.



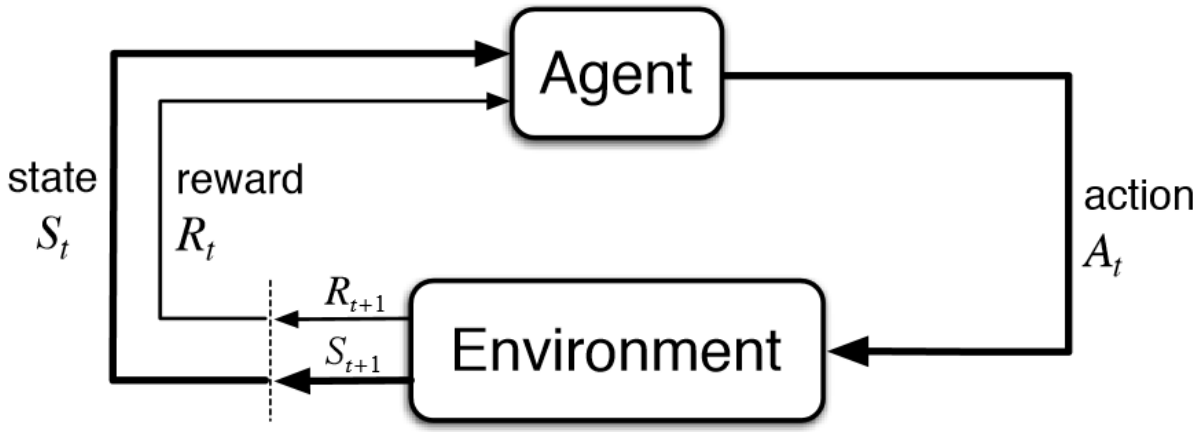
In the board game story, an 'episode' would be one full game. In this example, and in many RL problems, only the last state has a non-zero reward. **(source)** Research in RL has recently been reinvigorated by deep learning, but the basic model hasn't changed much; after all, this learning-from-scratch approach goes back to the very creation of RL as a research field and is encoded in its most fundamental equations.

So here's the basic question: **how reasonable is it to design AI models based on pure RL if pure RL makes so little intuitive sense?** If it's so absurd to conceive of a human learning a new board game through pure RL, shouldn't we wonder if it's a flawed framework for how AI agents should learn? Does it really make sense to start learning a new skill based only on its reward signal, with neither prior experience nor higher-level instruction?

Both prior experience and high-level instruction are completely missing from the equations classically used to formalize RL, and implicitly or explicitly altering those equations could have large implications for the algorithms we use to train AI algorithms for all applications of RL (which go well beyond board games, and range from robotics to resource allocation). In other words, this is a Big Question, and answering it will require two articles:

- In part one (this article), we'll begin by showing that the major accomplishments of pure RL are **not as impressive as they may seem**. Then, we'll go further by showing that more complex accomplishments may

not be possible under pure RL, due to the numerous **limitations it imposes** on AI agents.
- In part two, we'll overview the different approaches within AI that can address those limitations (chiefly, **meta-learning** and **zero-shot learning**). And finally, we'll get to a survey of monumentally exciting work based on these approaches, and conclude with what that work implies for the future of RL and AI as a whole.



Everyone agrees on pure RL as the basic formulation of reinforcement learning. But should they?

Does Pure RL Actually Make Sense?

Many people's immediate response goes something like the following:

Sure, it still makes sense to use pure RL — AI agents are not humans and do not have to learn like us, and pure RL has already been shown to solve all sorts of complex problems.

I disagree. By definition, AI research involves the ambition to enable machines to do things that only humans and animals are presently capable of. Therefore, comparison to human intelligence is appropriate. And as for the problems pure RL has been used to solve, an important caveat often goes unacknowledged: *those problems are generally not as complex as they may seem*.

This might be a surprise to many, since solving these problems has been the source of AI's most widely publicized accomplishments. While these are indeed great accomplishments, I nevertheless claim the problems involved are not as complex as they may seem. Before going into why that is the case, let us enumerate these accomplishments and point out why they are definitely worthy of praise:

- **DQN** — the research project by DeepMind that hugely increased interest in RL research just a brief 5 years ago by showing that combining deep learning with pure RL and a few new innovations could enable solving more complicated problems than ever before.
  It is not an exaggeration to say that DQN was the model that single-handedly revived researchers' interest in RL. Though it included only a few relatively simple innovations, those innovations proved hugely important for making 'Deep RL' practical.

Simple though it seems, learning to play this game from just the pixel inputs of the game would have been unthinkable just a decade ago.

- **AlphaGo Zero** and **AlphaZero** — the pure RL model that learned to play Go, Chess, and Shogi better than humanity's best.
  For those unaware, AlphaGo Zero is DeepMind's recent successor to AlphaGo (the program that first beat humanity's best at Go). Unlike the original AlphaGo, which learned through a combination of supervised learning and RL, AlphaGo Zero learns purely through RL and self-play. Thus, it follows the overall methodology of pure RL quite closely (with the agent starting with zero knowledge and learning from a reward signal), though it also uses a provided model (the rules of the game) and self-play to reliably and continuously get better.

DeepMind's own explanation of why AlphaGo Zero is so exciting.

Because it was no longer learning its success from humans, AlphaGo Zero was seen by many as even more of a game changer than AlphaGo. And then there was AlphaZero, a more generalized version that was shown to not only tackle Go but Chess and Shogi as well; this was the first time a single algorithm was used to crack both Chess and Go, and was not specifically tailored for either game like Deep Blue and the original AlphaGo were. For all the above reasons, AlphaGo Zero and AlphaZero are certainly monumental and exciting achievements (and great PR).

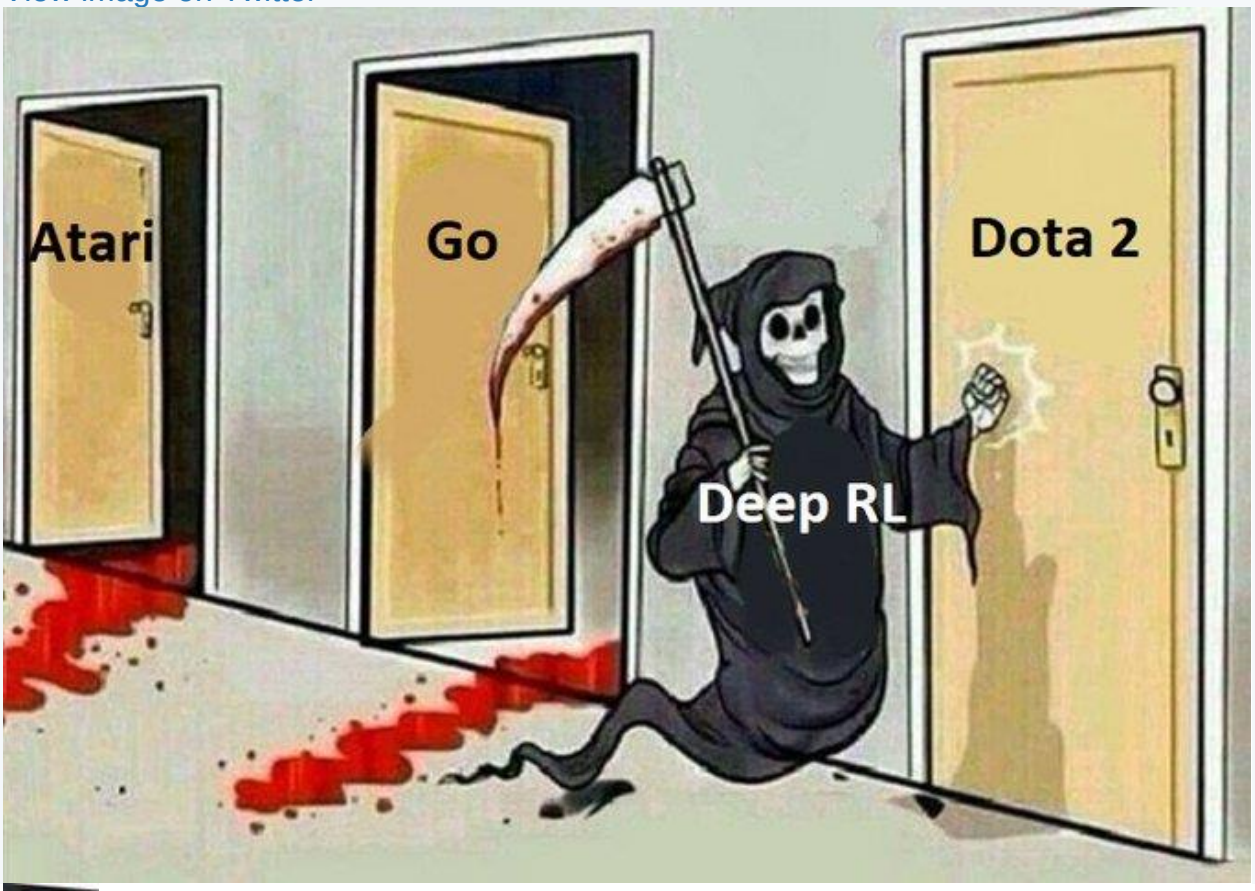
Lee Sedol losing a game to AlphaGo - a historic moment. caption>

- **OpenAI's Dota bots** — the Deep RL powered AI agents that can beat humans at the popular and complex competitive multiplayer game Dota 2. OpenAI's 2017 achievement of beating pros at a limited 1v1 variation of the game was impressive enough, but is nothing compared to their more recent feat of managing to beat a team of human players at a much more complex 5v5 variation of the game. It is also a successor to AlphaGo Zero in the sense that it also does not require any human knowledge and is trained purely through self-play. OpenAI themselves explain their achievement well:

There is no doubt doing well at this teamwork-based and highly complicated game is by far more impressive than the prior accomplishments of beating Atari games and Go. What is more, this was done without any major

algorithmic advance; the accomplishment was rather due to a truly astounding amount of computation and the use of an already well established pure RL algorithm as well as deep learning. Among the AI community, there was a common impression this was an impressive accomplishment and the next step in RL's string of huge milestones:

View image on Twitter



AI Memes for Artificially Intelligent Teens@ai_memes

Impressive work by @OpenAI, combining two of my favourite things, had to respond with my third favourite thing.

89
4:36 PM - Jun 25, 2018

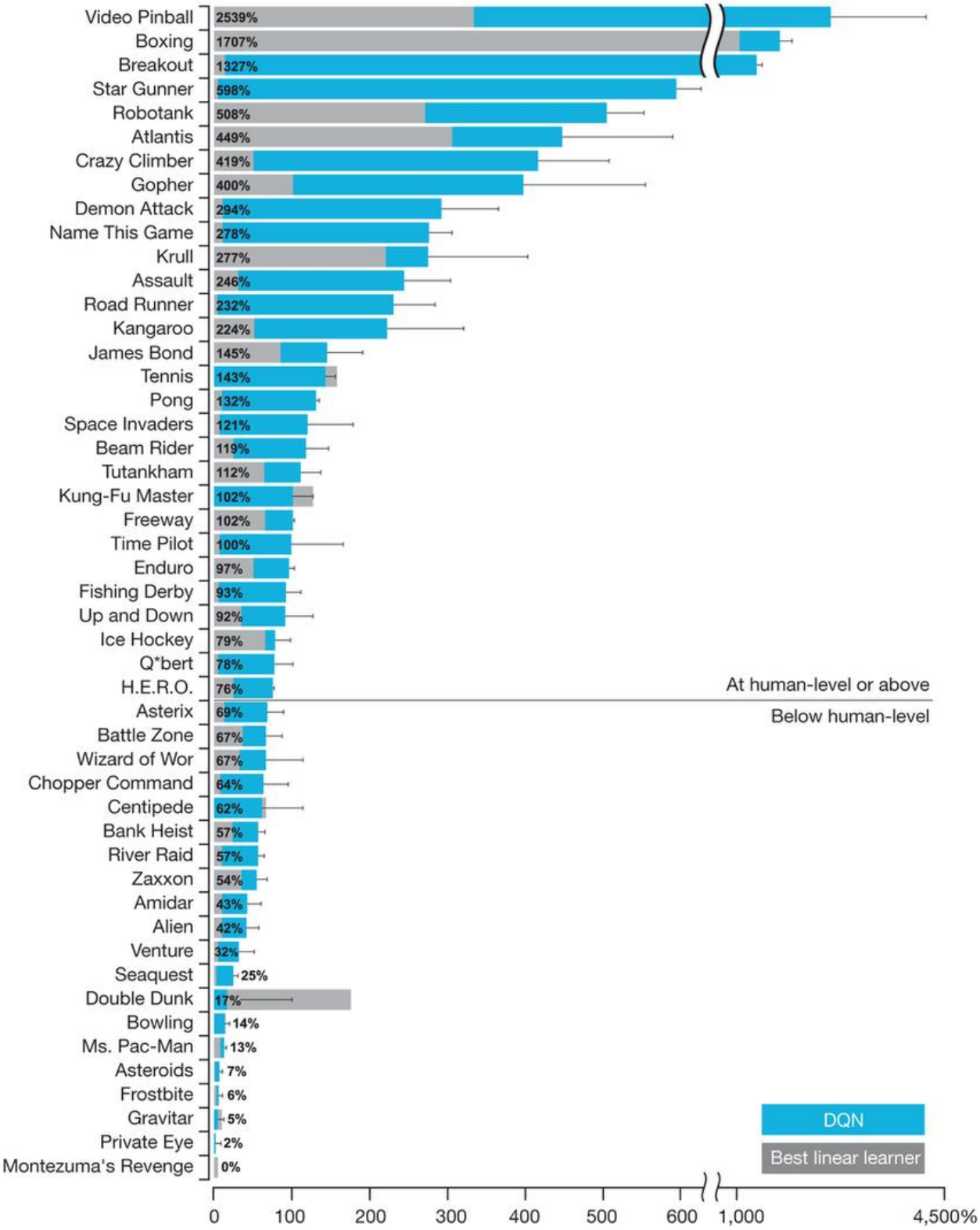
See AI Memes for Artificially Intelligent Teens's other Tweets

So yes, pure RL has achieved a lot. But let us now take a closer look and see why those achievements may not be as impressive as they seem.

On The Complexity of RL's Recent Successes

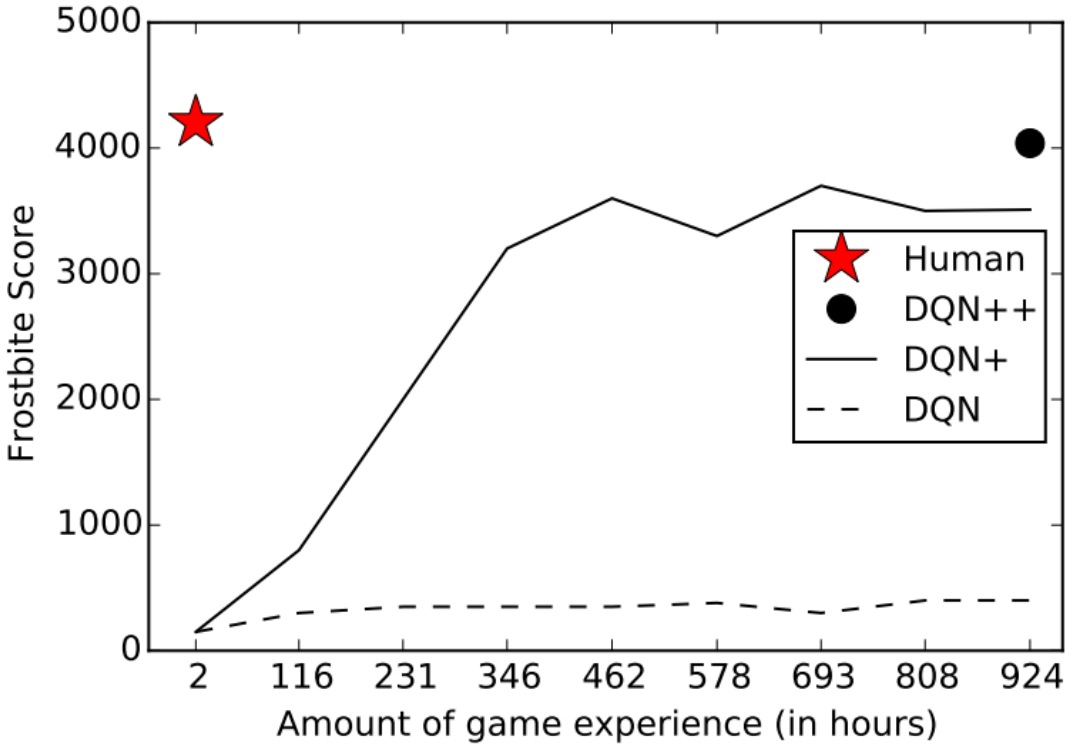
Let's start with DQN.

- It can achieve superhuman level play at many Atari games, but **far from all**. Generally, it is only able to do well at reflex-based games in which reasoning and memory are not required. Even 5 years later, no pure RL algorithms have cracked reasoning and memory games; on the contrary, approaches that have done well at them have either used instructions or demonstrations just as we mentioned would make sense to do in the board game allegory.



Though DQN is great at games like Breakout, it is still not able to tackle relatively simple games like Montezuma's Revenge. **(source)**

- Even for the game in which DQN can be supremely good, it requires **absurdly huge amounts of time and experience** to learn to do so compared to humans.



From **"Building Machines That Learn and Think Like People "**
The same limitations apply to AlphaGo Zero and AlphaZero. You see, Go is only hard within the context of the simplest category of AI problems. That is, it is in the category of problems with every property that makes a learning task easy: it is deterministic, discrete, static, fully observable, fully-known, single-agent, episodic, cheap and easy to simulate, easy to score... Only one thing is challenging about Go: its huge branching factor.

A Venn Diagram demonstration of Go's categorical simplicity.
So, Go might be the hardest easy problem, but it's still an easy problem. And predictions that AGI (Artificial General Intelligence) is imminent based only on AlphaGo's success can be safely dismissed — for all the reasons mentioned, most researchers recognize that the real world is vastly more complex than a simple game like Go. While impressive, all variations of AlphaGo are still fundamentally similar to Deep Blue: it's an expensive system, engineered over many years, with millions of dollars of investment, purely for the task of playing an abstract board game — and nothing else.

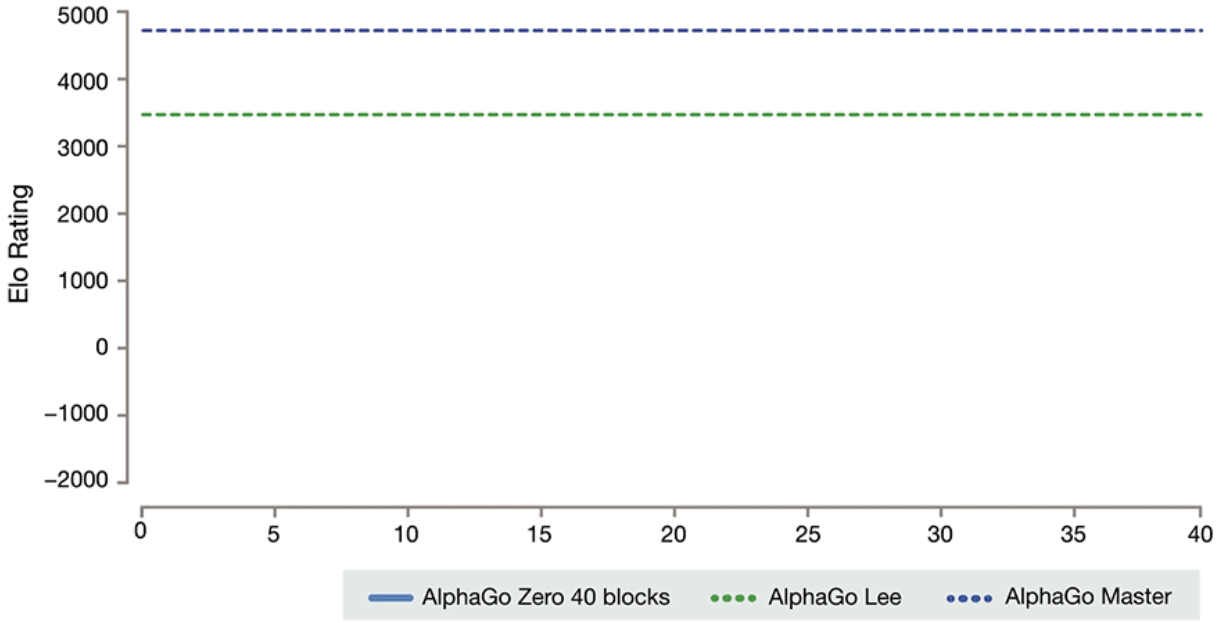
On to Dota. Yes, it is a far more complex game than Go and lacks many of the properties that make Go simple. It is not discrete, static, fully observable, single-agent, or episodic — a hugely challenging type of problem. But it is still an easily-simulated game controlled through a nifty API -- which entirely removes the need perception or motor control -- and so is ultimately simple compared to the true complexity of learning to solve problems in the real world as we do every day. And it is still like AlphaGo in that it took massive investment and many engineers to get an algorithm to solve the problem using an **absurdly huge amounts of time and experience** (it takes thousands of years' worth of game-playing experience to train and the use of a whopping 256 GPUs and 128,000 CPU cores).

So, despite definitely being huge accomplishments, there are some strong caveats to be aware of for all of them. Thus, arguing pure RL is just fine based solely on it having gotten us this far is not valid. And even all that aside, it must be considered — could pure RL just be the first but not best way to get to this accomplishments?

Pure RL's Fundamental Flaw - Starting from Scratch

Might there be a better way for AI agents to learn to play Go or Dota? The very name “AlphaGo Zero” is a reference to the idea that the model learns to play Go "from scratch". But let's recall that board game allegory. Trying to learn the board game 'from scratch' without explanation was absurd, right? So why is it a goal to strive towards with AI?

In fact, what if the board game you were trying to learn was Go — how would *you* start learning it? You would read the rules, learn some high-level strategies, recall how you played similar games in the past, get some advice... right? Indeed, it's at least partially because of the learning from scratch **limitation** of AlphaGo Zero and OpenAI's Dota bots that it is not truly impressive compared to human learning: they rely on seeing many orders of magnitude more games and using far more raw computational power than any human ever can.



The progression of AlphaGo Zero's skill. Note that it takes a whole day and thousands of lifetimes' worth of games to get to an ELO score of 0 (which even the weakest human can achieve easily). From **DeepMind's AlphaGo Zero Blog Post**

To be fair, pure RL techniques can be legitimately useful for 'narrow' tasks such as continuous control or more recently complex games such as Dota or Starcraft. However, with the success of deep learning, the AI research community as a whole is now trying to tackle ever more complex tasks that must deal with the limitless open-ended complexity of the real world (such as driving cars or holding conversations). It is for these less narrow tasks (that is, the majority of problems AI needs to tackle), and for the long term future of AI as a whole, that something beyond pure RL may be necessary.

So let's move on to tackling our revised question: is pure RL, and the idea of learning from scratch in general, the right approach for non-narrow/complex tasks?

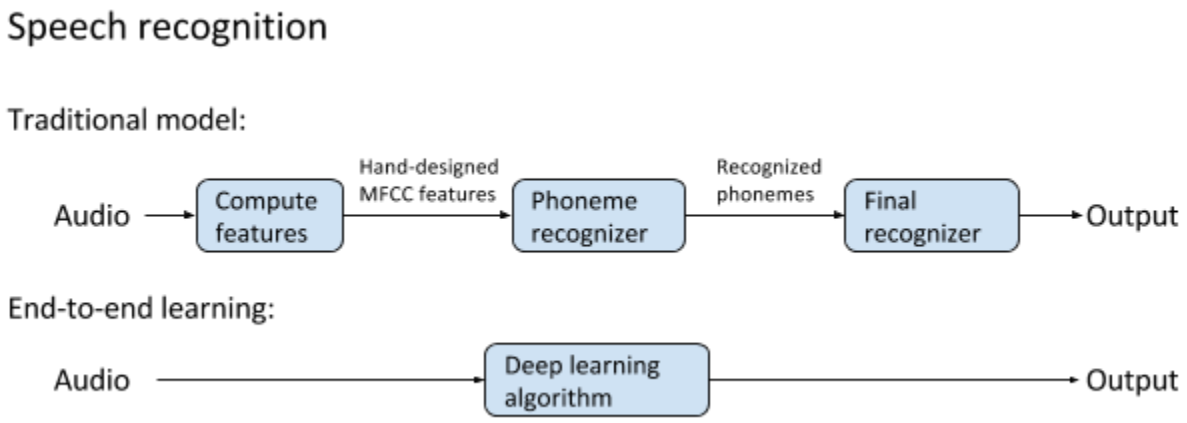
Should We Stick With Pure RL?

One answer to this question might be:

Yes, pure RL is the right approach to problems beyond those like Go or Dota. Though it makes no sense in the context of board games, it does make sense generally to learn things 'from scratch'. And, inspiration from humans aside, it makes sense to start from scratch so the agent

has no preconceptions and can be better than us (as with AlphaGo Zero).

Let's start with that last bit, ignoring human inspiration and considering the merits of learning from scratch in the context of AI in general. The typical justification of doing things “from scratch” is that the presumed alternative – hard-coding human intuitions into the model – might limit the model’s accuracy through unnecessary restrictions, or even worsen its performance with incorrect intuitions. This perspective has become mainstream with the success of deep learning methods, which learn 'end-to-end' models with millions of parameters, trained on staggering amounts of data and having only a few innate priors.



An illustration of both older non-traditional speech recognition and end-to-end deep learning methods. The latter works much better and is the basis for modern state of the art speech recognition. **(source)**
Here's the thing: incorporating prior knowledge or instructions doesn't necessitate imposing a lot of limiting structure based on human intuition on the learning agent. In other words, it is possible to inform a learning agent or model about the task at hand without limiting its ability to learn in the deep-learning style (that is, informed primarily by data, unlike Deep Blue and before that expert systems).

We'll get to concrete examples of techniques that let us do this soon, but the important point is that for most AI problems, not starting from scratch would not necessarily limit what the agent can learn in any way. There is no clear reason for an algorithm like AlphaGo Zero to emphasize starting from scratch so much when it can likely be bootstrapped with human knowledge (as was done with the original AlphaGo) or from learning other board games beforehand and still converge to the same superhuman level of skill. We'll get to concrete examples of techniques that do just that that soon...

Even if you care about none of this and just want to start from scratch, is pure RL the best way to do so? The answer used to be a no-brainer; in the domain of gradient-free optimization, pure RL was the most principled and trusted approach you could pick. But multiple recent papers have seriously questioned that stance by showing that the relatively simpler (and broadly less respected) evolution strategy-based methods seem to **do just about as well** on the same sorts of benchmarks pure RL has been routinely evaluated on:

- "Simple random search provides a competitive approach to reinforcement learning"
- "Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning"

Figure 2. **Example of high-performing individual on Frostbite found through random search**. See text for a description of the behavior of this policy. Its final score is 3,620 in this episode, which is higher than the scores produced by DQN, A3C and ES, although not as high as the score found by the GA (Table 1).

From **Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning**

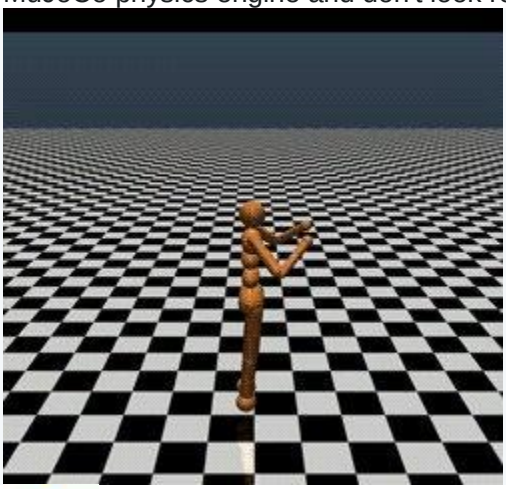
- "Evolution Strategies as a Scalable Alternative to Reinforcement Learning":

- "Towards Generalization and Simplicity in Continuous Control":

hardmaru@hardmaru

· Mar 20, 2018

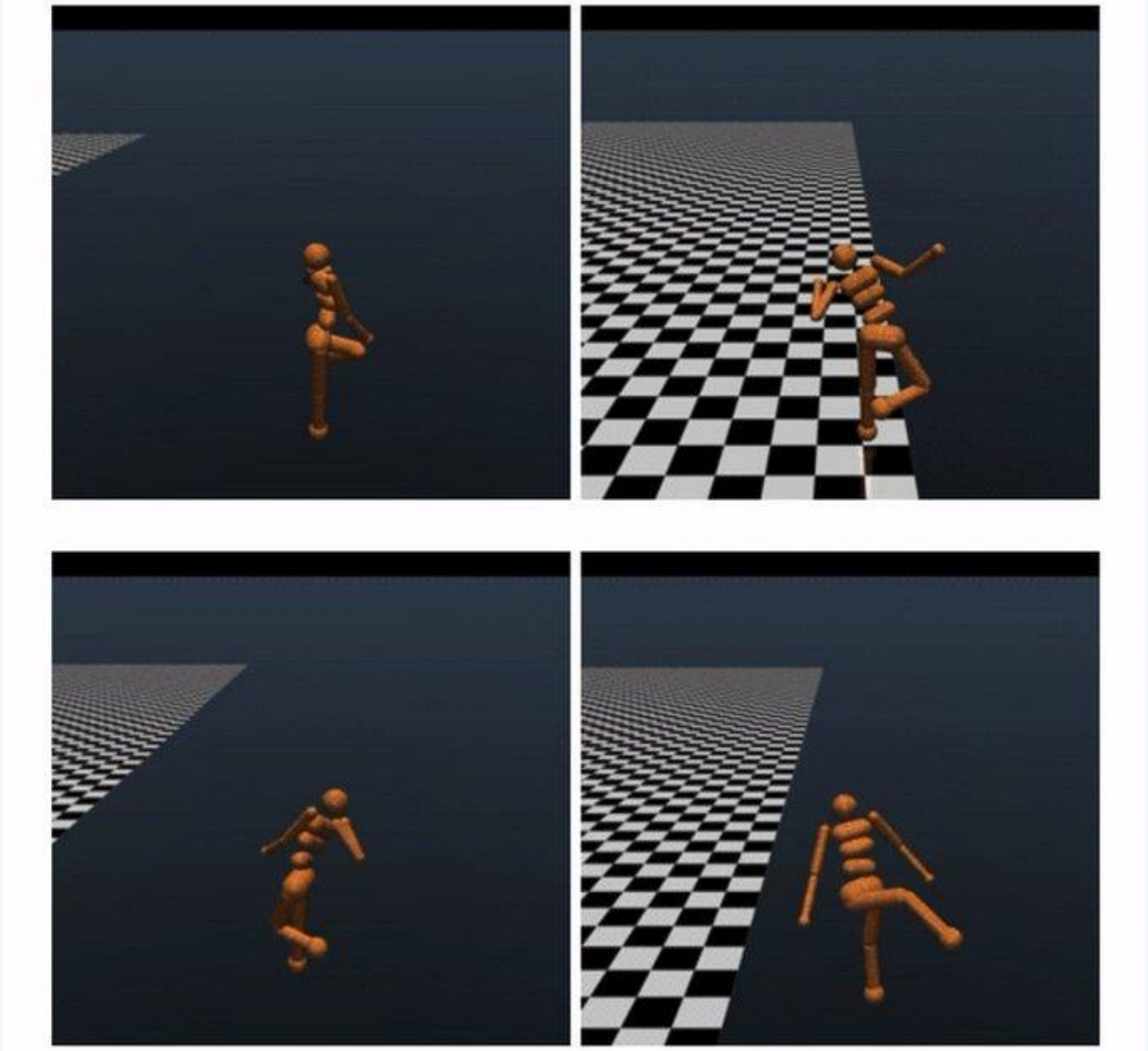
Replying to @hardmaru
Their Augmented Random Search method found solutions for MuJoCo Humanoid that obtained really high rewards (~11,600), but as they point out, these policies totally overfit to deficiencies of the MuJoCo physics engine and don't look realistic.



hardmaru@hardmaru

You may not like it, but this is what peak performance looks like.

503
1:18 PM - Mar 20, 2018
Twitter Ads info and privacy



179 people are talking about this

Ben Recht, a leading researcher on the theory and practice optimization algorithms and one of the authors of the "Simple random search provides a competitive approach to reinforcement learning" paper, has nicely summarized all the above points :

We have seen that random search works well on simple linear problems and appears better than some RL methods like policy gradient. But does random search break down as we move to harder problems? Spoiler Alert: No.

So, it is not clear pure RL is even the right way to do learning from scratch. But, let us get back to the question of human learning with regards to learning from scratch. Do people ever start learning a complex new skill (such as putting together new IKEA furniture or even driving a car) given no information (no information at all, not even prior experience) except for what possible actions they have as part of that skill? No, right?

Maybe for some very fundamental and general problems (such as the ones young babies deal with) it makes sense to start from scratch and do pure RL, since these problems are so broad it's hard to do anything else. But for the vast majority of problems in AI, there is no clear benefit in starting from scratch; we know what we want the AI agent to learn and can provide demonstrations or instructions for the skill. And in fact, starting from scratch is a primary reason for many of the widely agreed upon limitations of current AI and RL:

- Current AI is **data-hungry** (that is, sample-inefficient – in most cases, massive amounts of data are needed for state of the art AI methods to be useful. This is particularly bad for pure RL techniques; recall how AlphaGo Zero needed millions of games of Go to get to an ELO score of 0, which most people would manage after right away. By definition, learning from scratch is just about the least sample-efficient approach there can be.
- Current AI is **opaque** – in most cases, we have nothing but high-level intuitions about what an AI algorithm learns and how it works. For most AI problems, we want the algorithms to be predictable and explainable; a big neural net that just learns whatever it wants from scratch given just the low level reward signal and maybe an environment model (just how AlphaGo Zero works) is just about the least explainable and predictable approach to learning there can be.
- Current AI is **narrow** – in most cases, the AI models we build can only do one very narrow task and can easily be broken. Learning every single skill from scratch limits the ability to learn anything but one specific thing.
- Current AI is **brittle** – in most cases, our AI models generalize well to unseen inputs only with massive amounts of data and are even then still surprisingly easy to break.

So, we tend to know what we want the AI agent to learn. If the AI agent were a person, we could explain the task and probably provide some tips. But AI agents are not people -- might we do that for an AI agent? Turns out, in quite a few ways. Read on and find out in Part 2.

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

An expanded extract from this article, titled 'AlphaGo Zero Is Not A Sign of Imminent Human-Level AI' can be read on Skynet Today.