

Dota 2 with Large Scale Deep Reinforcement Learning

OpenAI, *

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung,
Przemysław "Psycho" Dębiak, Christy Dennison, David Farhi, Quirin Fischer,
Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson,
Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman,
Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang,
Filip Wolski, Susan Zhang

December 13, 2019

Abstract

On April 13th, 2019, OpenAI Five became the first AI system to defeat the world champions at an esports game. The game of Dota 2 presents novel challenges for AI systems such as long time horizons, imperfect information, and complex, continuous state-action spaces, all challenges which will become increasingly central to more capable AI systems. OpenAI Five leveraged existing reinforcement learning techniques, scaled to learn from batches of approximately 2 million frames every 2 seconds. We developed a distributed training system and tools for continual training which allowed us to train OpenAI Five for 10 months. By defeating the Dota 2 world champion (Team OG), OpenAI Five demonstrates that self-play reinforcement learning can achieve superhuman performance on a difficult task.

1 Introduction

The long-term goal of artificial intelligence is to solve advanced real-world challenges. Games have served as stepping stones along this path for decades, from Backgammon (1992) to Chess (1997) to Atari (2013)[1–3]. In 2016, AlphaGo defeated the world champion at Go using deep reinforcement learning and Monte Carlo tree search[4]. In recent years, reinforcement learning (RL) models have tackled tasks as varied as robotic manipulation[5], text summarization [6], and video games such as Starcraft[7] and Minecraft[8].

Relative to previous AI milestones like Chess or Go, complex video games start to capture the complexity and continuous nature of the real world. Dota 2 is a multiplayer real-time strategy game produced by Valve Corporation in 2013, which averaged between 500,000 and 1,000,000 concurrent players between 2013 and 2019. The game is actively played by full time professionals; the prize pool for the 2019 international championship exceeded \$35 million (the largest of any esports game in the world)[9, 10]. The game presents challenges for reinforcement learning due to long time horizons, partial observability, and high dimensionality of observation and action spaces. Dota 2's

*Authors listed alphabetically. Please cite as OpenAI et al., and use the following bibtex for citation: <https://openai.com/bibtex/openai2019dota.bib>

rules are also complex — the game has been actively developed for over a decade, with game logic implemented in hundreds of thousands of lines of code.

The key ingredient in solving this complex environment was to scale existing reinforcement learning systems to unprecedented levels, utilizing thousands of GPUs over multiple months. We built a distributed training system to do this which we used to train a Dota 2-playing agent called OpenAI Five. In April 2019, OpenAI Five defeated the Dota 2 world champions (Team OG¹), the first time an AI system has beaten an esports world champion². We also opened OpenAI Five to the Dota 2 community for competitive play; OpenAI Five won 99.4% of over 7000 games.

One challenge we faced in training was that the environment and code continually changed as our project progressed. In order to train without restarting from the beginning after each change, we developed a collection of tools to resume training with minimal loss in performance which we call *surgery*. Over the 10-month training process, we performed approximately one surgery per two weeks. These tools allowed us to make frequent improvements to our strongest agent within a shorter time than the typical practice of training from scratch would allow. As AI systems tackle larger and harder problems, further investigation of settings with ever-changing environments and iterative development will be critical.

In section 2, we describe Dota 2 in more detail along with the challenges it presents. In section 3 we discuss the technical components of the training system, leaving most of the details to appendices cited therein. In section 4, we summarize our long-running experiment and the path that lead to defeating the world champions. We also describe lessons we’ve learned about reinforcement learning which may generalize to other complex tasks.

2 Dota 2

Dota 2 is played on a square map with two teams defending bases in opposite corners. Each team’s base contains a structure called an ancient; the game ends when one of these ancients is destroyed by the opposing team. Teams have five players, each controlling a hero unit with unique abilities. During the game, both teams have a constant stream of small “creep” units, uncontrolled by the players, which walk towards the enemy base attacking any opponent units or buildings. Players gather resources such as gold from creeps, which they use to increase their hero’s power by purchasing items and improving abilities.³

To play Dota 2, an AI system must address various challenges:

- **Long time horizons.** Dota 2 games run at 30 frames per second for approximately 45 minutes. OpenAI Five selects an action every fourth frame, yielding approximately 20,000 steps per episode. By comparison, chess usually lasts 80 moves, Go 150 moves[11].
- **Partially-observed state.** Each team in the game can only see the portion of the game state near their units and buildings; the rest of the map is hidden. Strong play requires making inferences based on incomplete data, and modeling the opponent’s behavior.

¹<https://www.facebook.com/OGDota2/>

²Full game replays and other supplemental can be downloaded from: <https://openai.com/blog/how-to-train-your-openai-five/>

³Further information the rules and gameplay of Dota 2 is readily accessible online; a good introductory resource is <https://purgegamers.true.io/g/dota-2-guide/>

- **High-dimensional action and observation spaces.** Dota 2 is played on a large map containing ten heroes, dozens of buildings, dozens of non-player units, and a long tail of game features such as runes, trees, and wards. OpenAI Five observes $\sim 16,000$ total values (mostly floats and categorical values with hundreds of possibilities) each time step. We discretize the action space; on an average timestep our model chooses among 8,000 to 80,000 actions (depending on hero). For comparison Chess requires around one thousand values per observation (mostly 6-possibility categorical values) and Go around six thousand values (all binary)[12]. Chess has a branching factor of around 35 valid actions, and Go around 250[11].

Our system played Dota 2 with two limitations from the regular game:

- Subset of 17 heroes — in the normal game players select before the game one from a pool of 117 heroes to play; we support 17 of them.⁴
- No support for items which allow a player to temporarily control multiple units at the same time (Illusion Rune, Helm of the Dominator, Manta Style, and Necronomicon). We removed these to avoid the added technical complexity of enabling the agent to control multiple units.

3 Training System

3.1 Playing Dota using AI

Humans interact with the Dota 2 game using a keyboard, mouse, and computer monitor. They make decisions in real time, reason about long-term consequences of their actions, and more. We adopt the following framework to translate the vague problem of “play this complex game at a superhuman level” into a detailed objective suitable for optimization.

Although the Dota 2 engine runs at 30 frames per second, OpenAI Five only acts on every 4th frame which we call a *timestep*. Each timestep, OpenAI Five receives an *observation* from the game engine encoding all the information a human player would see such as units’ health, position, etc (see Appendix E for an in-depth discussion of the observation). OpenAI Five then returns a discrete *action* to the game engine, encoding a desired movement, attack, etc.

Certain game mechanics were controlled by hand-scripted logic rather than the policy: the order in which heroes purchase items and abilities, control of the unique courier unit, and which items heroes keep in reserve. While we believe the agent could ultimately perform better if these actions were not scripted, we achieved superhuman performance before doing so. Full details of our action space and scripted actions are described in Appendix F.

Some properties of the environment were randomized during training, including the heroes in the game and which items the heroes purchased. Sufficiently diverse training games are necessary to ensure robustness to the wide variety of strategies and situations that arise in games against human opponents. See subsection O.2 for details of the domain randomizations.

We define a *policy* (π) as a function from the history of observations to a probability distribution over actions, which we parameterize as a recurrent neural network with approximately 159 million parameters (θ). The neural network consists primarily of a single-layer 4096-unit LSTM [13] (see Figure 1). Given a policy, we play games by repeatedly passing the current observation as input and sampling an action from the output distribution at each timestep.

⁴See Appendix P for experiments characterizing the effect of hero pool size.

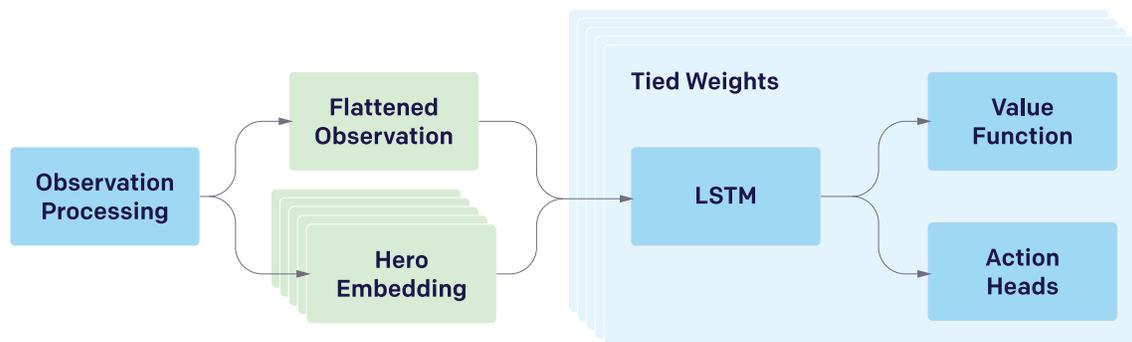


Figure 1: **Simplified OpenAI Five Model Architecture:** The complex multi-array observation space is processed into a single vector, which is then passed through a 4096-unit LSTM. The LSTM state is projected to obtain the policy outputs (actions and value function). Each of the five heroes on the team is controlled by a replica of this network with nearly identical inputs, each with its own hidden state. The networks take different actions due to a part of the observation processing’s output indicating which of the five heroes is being controlled. The LSTM composes 84% of the model’s total parameter count. See Figure 17 and Figure 18 in Appendix H for a detailed breakdown of our model architecture.

Separate replicas of the same policy function (with identical parameters θ) are used to control each of the five heroes on the team. Because visible information and *fog of war* (area that is visible to players due to proximity of friendly units) are shared across a team in Dota 2, the observations are nearly⁵ identical for each hero.

Instead of using the pixels on the screen, we approximate the information available to a human player in a set of data arrays (see Appendix E for full details of the observations space). This approximation is imperfect; there are small pieces of information which humans can gain access to which we have not encoded in the observations. On the flip side, while we were careful to ensure that all the information available to the model is also available to a human, the model does get to see *all* the information available simultaneously every time step, whereas a human needs to actively click to see various parts of the map and status modifiers. OpenAI Five uses this semantic observation space for two reasons: First, because our goal is to study strategic planning and high-level decision-making rather than focus on visual processing. Second, it is infeasible for us to render each frame to pixels in all training games; this would multiply the computation resources required for the project many-fold. Although these discrepancies exist, we do not believe they introduce significant bias when benchmarking against human players. To allow the five networks to choose different actions, the LSTM receives an extra input from the observation processing, indicating which of the five heroes is being controlled, detailed in Figure 17.

Because of the expansive nature of the problem and the size and expense of each experiment, it was not practical to investigate all the details of the policy and training system. Many details, even

⁵We do include a very small number of derived features which depend on the hero being controlled, for example the “distance to me” feature of each unit in the game.

some large ones, were set for historical reasons or on the basis of preliminary investigations without full ablations.

3.2 Optimizing the Policy

Our goal is to find a policy which maximizes the probability of winning the game against professional human experts. In practice, we maximize a *reward function* which includes additional signals such as characters dying, collecting resources, etc. We also apply several techniques to exploit the zero-sum multiplayer structure of the problem when computing the reward function — for example, we symmetrize rewards by subtracting the reward earned by the opposing team. We discuss the details of the reward function in Appendix G. We constructed the reward function once at the start of the project based on team members’ familiarity with the game. Although we made minor tweaks when game versions changed, we found that our initial choice of what to reward worked fairly well. The presence of these additional signals was important for successful training (as discussed in Appendix G).

The policy is trained using Proximal Policy Optimization (PPO)[14], a variant of advantage actor critic[15, 16].⁶ The optimization algorithm uses Generalized Advantage Estimation [17] (GAE), a standard advantage-based variance reduction technique [15] to stabilize and accelerate training. We train a network with a central, shared LSTM block, that feeds into separate fully connected layers producing policy and value function outputs.

The training system is represented in Figure 2. We train our policy using collected self-play experience from playing Dota 2, similar to [18]. A central pool of *optimizer GPUs* receives game data and stores it asynchronously in local buffers called *experience buffers*. Each optimizer GPU computes gradients using minibatches sampled randomly from its experience buffer. Gradients are averaged across the pool using NCCL2[19] allreduce before being synchronously applied to the parameters. In this way the effective batch size is the batch size on each GPU (120 samples, each with 16 timesteps) multiplied by the number of GPUs (up to 1536 at the peak), for a total batch size of 2,949,120 time steps (each with five hero policy replicas).

We apply the Adam optimizer [20] using truncated backpropagation through time[21] over samples of 16 timesteps. Gradients are additionally clipped per parameter to be within between $\pm 5\sqrt{v}$ where v is the running estimate of the second moment of the (unclipped) gradient. Every 32 gradient steps, the optimizers publish a new *version* of the parameters to a central Redis⁷ storage called the *controller*. The controller also stores all metadata about the state of the system, for stopping and restarting training runs.

“Rollout” worker machines run self-play games. They run these games at approximately 1/2 real time, because we found that we could run slightly more than twice as many games in parallel at this speed, increasing total throughput. We describe our integration with the Dota 2 engine in Appendix K. They play the latest policy against itself for 80% of games, and play against older policies for 20% of games (for details of opponent sampling, see Appendix N). The rollout machines run the game engine but not the policy; they communicate with a separate pool of GPU machines which run forward passes in larger batches of approximately 60. These machines frequently poll the controller to gather the newest parameters.

⁶Early on in the project, we considered other algorithms including other policy gradient methods, q-learning, and evolutionary strategies. PPO was the first to show initial learning progress.

⁷<http://redis.io>

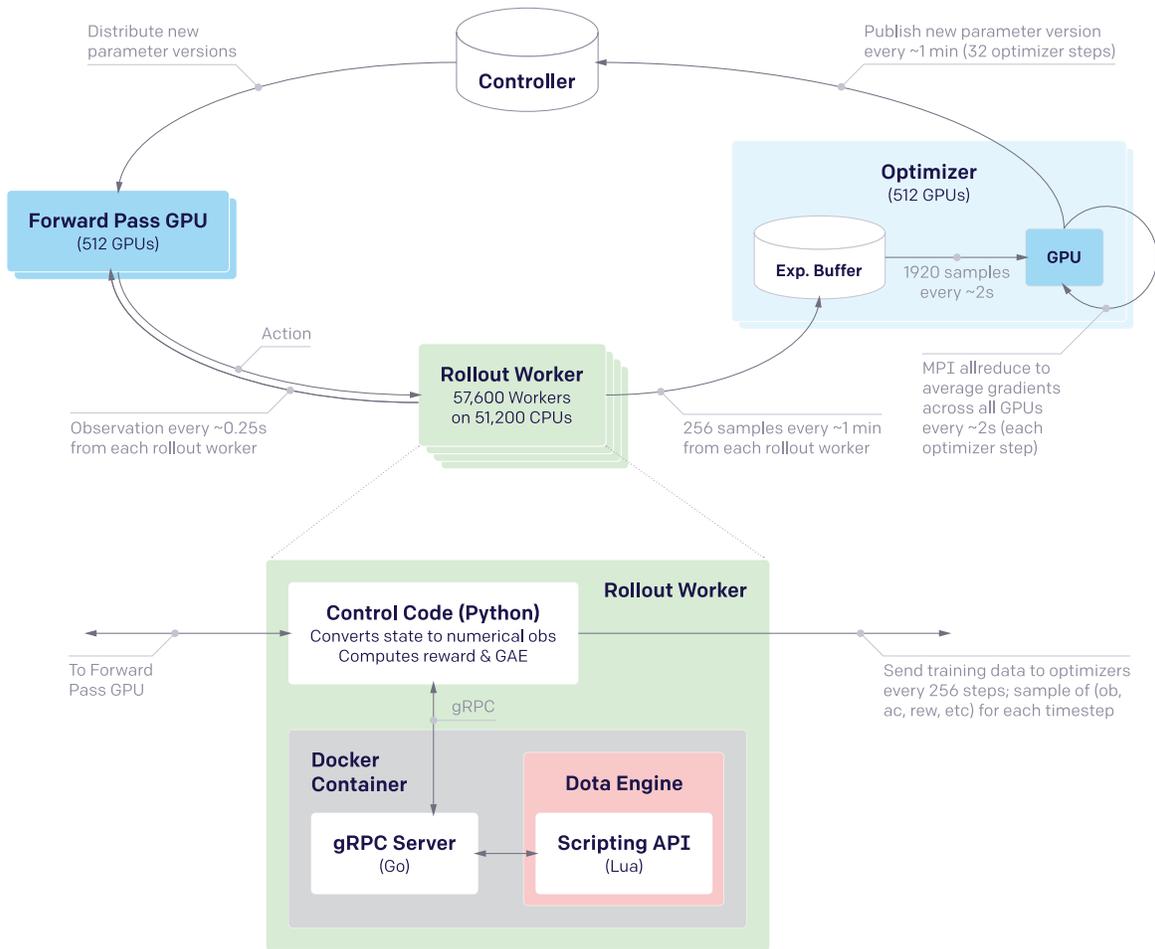


Figure 2: **System Overview**: Our training system consists of 4 primary types of machines. Rollouts run the Dota 2 game on CPUs. They communicate in a tight loop with Forward Pass GPUs, which sample actions from the policy given the current observation. Rollouts send their data to Optimizer GPUs, which perform gradient updates. The Optimizers publish the parameter versions to storage in the Controller, and the Forward Pass GPUs occasionally pull the latest parameter version. Machine numbers are for the Rerun experiment described in subsection 4.2; OpenAI Five’s numbers fluctuated between this scale and approximately 3x larger.

Rollout machines send data asynchronously from games that are in progress, instead of waiting for an entire game to finish before publishing data for optimization⁸; see Figure 8 in Appendix C for more discussion of how rollout data is aggregated. See Figure 5b for the benefits of keeping the rollout-optimization loop tight. Because we use GAE with $\lambda = 0.95$, the GAE rewards need to be smoothed over a number of timesteps $\gg 1/\lambda = 20$; using 256 timesteps causes relatively little loss.

The entire system runs on our custom distributed training platform called Rapid[5], running on Google Cloud Platform. We use ops from the blocksparse library for fast GPU training[22]. For a full list of the hyperparameters used in training, see Appendix C.

3.3 Continual Transfer via Surgery

As the project progressed, our code and environment gradually changed for three different reasons:

1. As we experimented and learned, we implemented changes to the training process (reward structure, observations, etc) or even to the architecture of the policy neural network.
2. Over time we expanded the set of game mechanics supported by the agent’s action and observation spaces. These were not introduced gradually in an effort to build a perfect curriculum. Rather they were added incrementally as a consequence of following the standard engineering practice of building a system by starting simple and adding complexity piece by piece over time.
3. From time to time, Valve publishes a new Dota 2 version including changes to the core game mechanics and the properties of heroes, items, maps, etc; to compare to human players our agent must play on the latest game version.

These changes can modify the shapes and sizes of the model’s layers, the semantic meaning of categorical observation values, etc.

When these changes occur, most aspects of the old model are likely relevant in the new environment. But cherry-picking parts of the parameter vector to carry over is challenging and limits reproducibility. For these reasons training from scratch is the safe and common response to such changes.

However, training OpenAI Five was a multi-month process with high capital expenditure, motivating the need for methods that can persist models across domain and feature changes. It would have been prohibitive (in time and money) to train a fresh model to a high level of skill after each such change (approximately every two weeks). For example, we changed to Dota 2 version 7.21d, eight days before our match against the world champions (OG); this would not have been possible if we had not continued from the previous agent.

Our approach, which we term “surgery”, can be viewed as a collection of tools to perform offline operations to the old model π_θ to obtain a new model $\hat{\pi}_{\hat{\theta}}$ compatible with the new environment, which performs at the same level of skill even if the parameter vectors $\hat{\theta}$ and θ have different sizes and semantics. We then begin training in the new environment using $\hat{\pi}_{\hat{\theta}}$. In the simplest case where the environment, observation, and action spaces did not change, our standard reduces to insisting

⁸Rollout machines produce 7.5 steps per second; they send data every 256 steps, or 34 seconds of game play. Because our rollout games run at approximately half-speed, this means they push data approximately once per minute.

that the new policy implements the same function from observed states to action probabilities as the old:

$$\forall o \hat{\pi}_{\hat{\theta}}(o) = \pi_{\theta}(o) \tag{1}$$

This case is a special case of *Net2Net*-style function preserving transformations [23]. We have developed tools to implement Equation 1 exactly when possible (adding observations, expanding layers, and other situations), and approximately when the type of modification to the environment, observation space, or action space precludes satisfying it exactly. See Appendix B for further discussion of surgery.

In the end, we performed over twenty surgeries (along with many unsuccessful surgery attempts) over the ten-month lifetime of OpenAI Five (see Table 1 in Appendix B for a full list). Surgery enabled continuous training without loss in performance (see Figure 4). In subsection 4.2 we discuss our experimental verification of this method.

4 Experiments and Evaluation

OpenAI Five is a single training run that ran from June 30th, 2018 to April 22nd, 2019. After ten months of training using 770 ± 50 PFlops/s-days of compute, it defeated the Dota 2 world champions in a best-of-three match and 99.4% of human players during a multi-day online showcase.

In order to utilize this level of compute effectively we had to scale up along three axes. First, we used batch sizes of 1 to 3 million timesteps (grouped in unrolled LSTM windows of length 16). Second, we used a model with over 150 million parameters. Finally, OpenAI Five trained for 180 days (spread over 10 months of real time due to restarts and reverts). Compared AlphaGo[4], we use 50 to 150 times larger batch size, 20 times larger model, and 25 times longer training time. Simultaneous works in recent months[7, 24] have matched or slightly exceeded our scale.

4.1 Human Evaluation

Over the course of training, OpenAI Five played games against numerous amateur players, professional players, and professional teams in order to gauge progress. For a complete list of the professional teams OpenAI Five played against over time, see Appendix I.

On April 13th, OpenAI Five played a high-profile game against OG, the reigning Dota 2 world champions, winning a best-of-three (2-0) and demonstrating that our system can learn to play at the highest levels of skill. For detailed analysis of our agent’s performance during this game and its overall understanding of the environment, see Appendix D.

Machine Learning systems often behave poorly when confronted with unexpected situations[25]. While winning a single high-stakes showmatch against the world champion indicates a very high level of skill, it does not prove a broad understanding of the variety of challenges the human community can present. To explore whether OpenAI Five could be consistently exploited by creative or out-of-distribution play, we ran *OpenAI Five Arena*, in which we opened OpenAI Five to the public for competitive online games from April 18-21, 2019. In total, Five played 3,193 teams in 7,257 total games, winning 99.4%⁹. Twenty-nine teams managed to defeat OpenAI Five for a total of 42 games

⁹Human players often abandoned losing games rather than playing them to the end, even abandoning games right after an unfavorable hero selection draft before the main game begins. OpenAI Five does not abandon games, so we count abandoned games as wins for OpenAI Five. These abandoned games (3140 of the 7215 wins) likely includes a small number of games that were abandoned for technical or personal reasons.

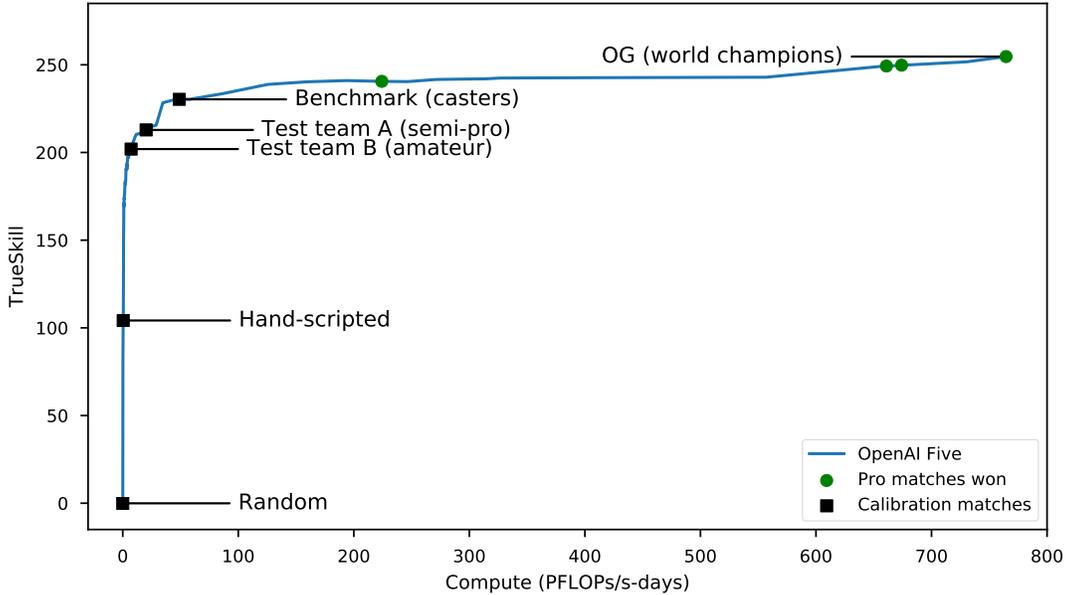


Figure 3: TrueSkill over the course of training for OpenAI Five. To provide informal context for how TrueSkill corresponds to human skill, we mark the level at which OpenAI Five begins to defeat various opponents, from random to world champions. Note that this is biased against earlier models; this TrueSkill evaluation is performed using the final policy and environment (Dota 2 version 7.21d, all non-illusion items, etc), even though earlier models were trained in the earlier environment. We believe this contributes to the inflection point around 600 PFLOPs/s-days — around that point we gave the policy control of a new action (buyback) and performed a major Dota 2 version upgrade (7.20). We speculate that the rapid increase to TrueSkill 200 early in training is due to the exponential nature of the scale — a constant TrueSkill difference of approximately 8.3 corresponds to an 80% winrate, and it is easier to learn how to consistently defeat bad agents.

lost.

In Dota 2, the key measure of human dexterity is *reaction time*¹⁰. OpenAI Five can react to a game event in 217ms on average. This quantity does not vary depending on game state. It is difficult to find reliable data on Dota 2 professionals’ reaction times, but typical human visual reaction time is approximately 250ms[26]. See Appendix L for more details.

While human evaluation is the ultimate goal, we also need to evaluate our agents continually during training in an automated way. We achieve this by comparing them to a pool of fixed reference agents with known skill using the TrueSkill rating system [27]. In our TrueSkill environment, a rating of 0 corresponds to a random agent, and a difference of approximately 8.3 TrueSkill between two agents roughly corresponds to an 80% winrate of one versus the other (see Appendix J for details of our TrueSkill setup). OpenAI Five’s TrueSkill rating over time can be seen in Figure 3.

OpenAI Five’s “playstyle” is difficult to analyze rigorously (and is likely influenced by our shaped reward function) but we can discuss in broad terms the flavor of comments human players made to describe how our agent approached the game. Over the course of training, OpenAI Five developed

¹⁰Contrast with RTS games like Starcraft, where the key measure is actions per minute due to the large number of units that need to be supplied with actions.

a distinct style of play with noticeable similarities and differences to human playstyles. Early in training, OpenAI Five prioritized large group fights in the game as opposed to accumulating resources for later, which led to games where they were significantly behind if the enemy team avoided fights early. This playstyle was risky and would result in quick wins in under 20 minutes if OpenAI Five got an early advantage, but had no way to recover from falling behind, leading to long and drawn out losses often over 45 minutes.

As the agents improved, the playstyle evolved to align closer with human play while still maintaining many of the characteristics learned early on. OpenAI Five began to concentrate resources in the hands of its strongest heroes, which is common in human play. Five relied heavily on large group battles, effectively applying pressure when holding a significant advantage, but also avoided fights and focused on gathering resources if behind.

The final agent played similar to humans in many broad areas, but had a few interesting differences. Human players tend to assign heroes to different areas of the map and only reassign occasionally, but OpenAI Five moved heroes back and forth across the map much more frequently. Human players are often cautious when their hero has low health; OpenAI Five seemed to have a very finely-tuned understanding of when an aggressive attack with a low-health hero was worth a risk. Finally OpenAI Five tended to more readily consume resources, as well as abilities with long *cooldowns* (time it takes to reload), while humans tend to hold on to those in case a better opportunity arises later.

4.2 Validating Surgery with Rerun

In order to validate the time and resources saved by our surgery method (see subsection 3.3), we trained a second agent between May 18, 2019 and June 12, 2019, using only the final environment, model architecture, etc. This training run, called “Rerun”, did not go through a tortuous route of changing game rules, modifications to the neural network parameters, online experiments with hyperparameters, etc.

Rerun took 2 months and 150 ± 5 PFlops/s-days of compute (see Figure 4). This timeframe is significantly longer than the frequency of our surgery changes (which happened every 1-2 weeks). As a naive comparison, if we had trained from scratch after each of our twenty major surgeries, the project would have taken 40 months instead of 10 (in practice we likely would have made fewer changes). Another benefit of surgery was that we had a very high-skill agent available for evaluation at all times, significantly tightening the iteration loop for experimental changes. In OpenAI Five’s regime — exploring a novel task and building a novel environment — perpetual training is a significant benefit.

Of course, in situations where the environment is pre-built and well-understood from the start, we see little need for surgery. Rerun took approximately 20% of the resources of OpenAI Five; if we had access to the final training environment ahead of time there would be no reason to start training e.g. on a different version of the game.

Rerun continued to improve beyond OpenAI Five’s skill, and reached over 98% winrate against the final version of OpenAI Five. We wanted to validate that our final code and hyperparameters would reproduce OpenAI Five performance, so we ceased training at that point. We believe Rerun would have continued improving, both because of its upward trend and because we had yet to fully anneal hyperparameters like learning rate and horizon to their final OpenAI Five settings.

This process of surgery successfully allowed us to change the environment every week. However, the model ultimately plateaued at a weaker skill level than the from-scratch model was able to

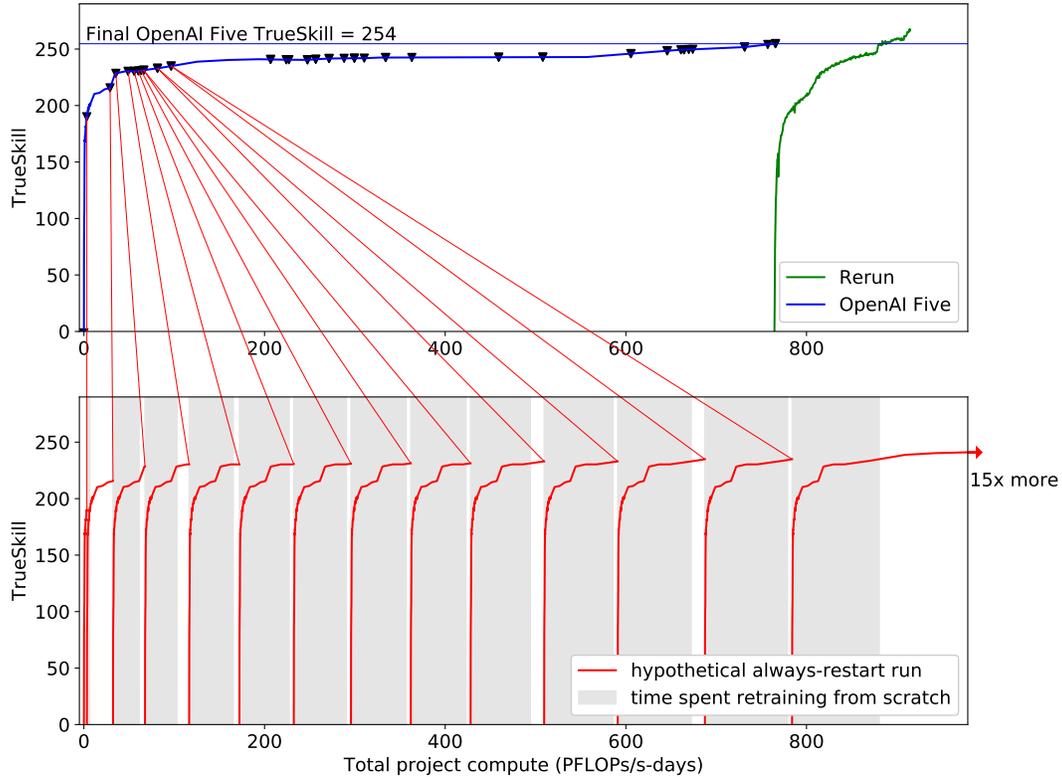


Figure 4: **Training in an environment under development:** In the top panel we see the full history of our project - we used surgery methods to continue training OpenAI Five at each environment or policy change without loss in performance; then we restarted once at the end to run Rerun. On the bottom we see the hypothetical alternative, if we had restarted after each change and waited for the model to reach the same level of skill (assuming pessimistically that the curve would be identical to OpenAI Five). The ideal option would be to run Rerun-like training from the very start, but this is impossible — the OpenAI Five curve represents lessons learned that led to the final codebase, environment, etc., without which it would not be possible to train Rerun.

achieve. Learning how to continue long-running training without affecting final performance is a promising area for future work.

Ultimately, while surgery as currently conceived is far from perfect, with proper tooling it becomes a useful method for incorporating certain changes into long-running experiments without paying the cost of a restart for each.

4.3 Batch Size

In this section, we evaluate the benefits of increasing the batch size using small scale experiments. Increasing the batch size in our case means two things: first, using twice as many optimizer GPUs to optimize over the larger batch, and second, using twice as many rollout machines and forward pass GPUs to produce twice as many samples to feed the increased optimizer pool.

One compelling benchmark to compare against when increasing the batch size is *linear* speedup: using 2x as much compute gets to the same skill level in 1/2 the time. If this scaling property holds, it is possible to use the same *total* amount of GPU-days (and thus dollars) to reach a given result[28]. In practice we see less than this ideal speedup, but the speedup from increasing batch size is still noticeable and allows us to reach the result in less wall time.

To understand how batch size affects training speed, we calculate the “speedup” of an experiment to reach various TrueSkill thresholds, defined as:

$$\text{speedup}(T) = \frac{\text{Versions for baseline to first reach TrueSkill } T}{\text{Versions for experiment to first reach TrueSkill } T} \tag{2}$$

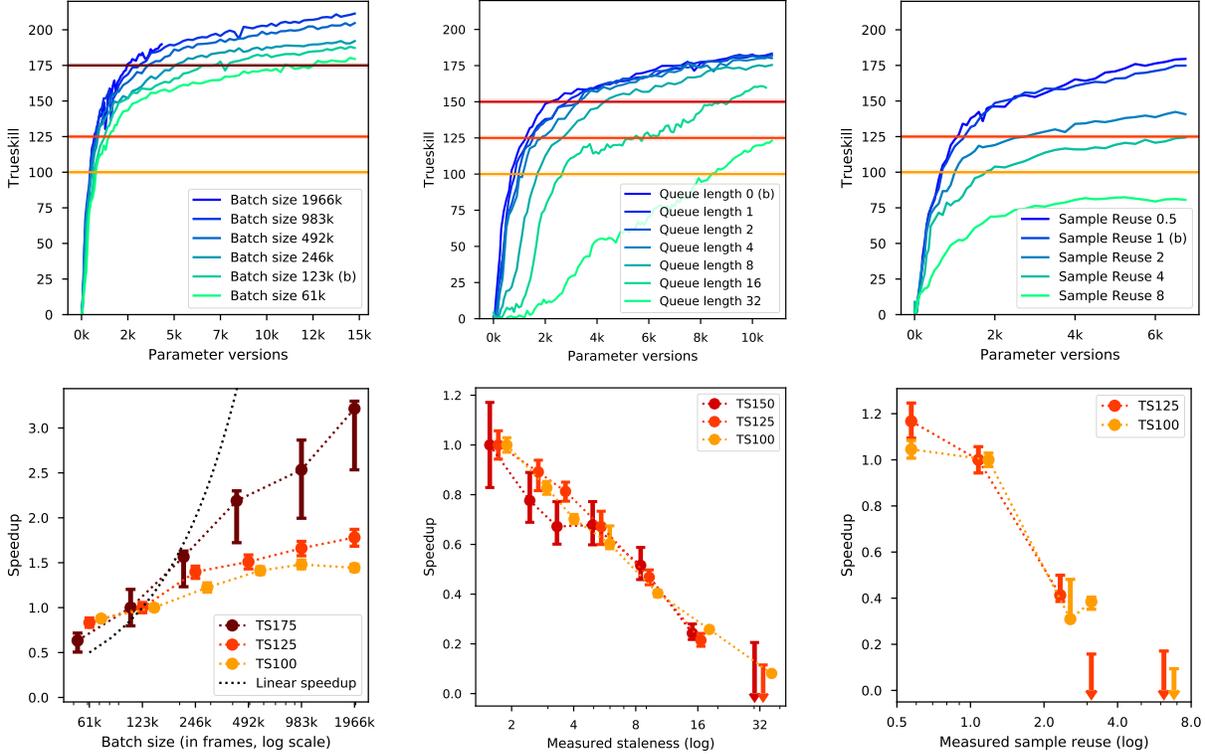
The results of varying batch size in the early part of training can be seen in Figure 5. Full details of the experimental setup can be found in Appendix M. We find that increasing the batch size speeds up training through the regime we tested, up to batches of millions of observations.

Using the scale of Rerun, we were able to reach superhuman performance in two months. In Figure 5a, we see that Rerun’s batch size (983k time steps) had a speedup factor of around 2.5x over the baseline batch size (123k). If we had instead used the smaller batch size, then, we might expect to wait 5 months for the same result. We speculate that it would likely be longer, as the speedup factor of 2.5 applies at TrueSkill 175 early in training, but it appears to increase with higher TrueSkill.

Per results in [28], we hoped to find (in the early part of training) linear speedup from increasing batch size; i.e. that it would be 2x faster to train an agent to certain thresholds if we use 2x the compute and data. Our results suggest that speedup is less than linear. However, we speculate that this may change later in training when the problem becomes more difficult. Also, given the relevant compute costs, in this ablation study we did not tune hyperparameters such as learning rate separately for each batch size.

4.4 Data Quality

One unusual feature of our task is the length of the games; each rollout can take up to two hours to complete. For this reason it is infeasible for us to optimize entirely on fully *on-policy* trajectories; if we waited to apply gradient updates for an entire rollout game to be played using the latest parameters, we could make only one update every two hours. Instead, our rollout workers and optimizers operate asynchronously: rollout workers download the latest parameters, play a small



(a) **Batch size:** Larger batch size speeds up training. In the early part of training studied here, the speedup is sublinear in the computation and samples required. See subsection M.1 for experiment details.

(b) **Data Staleness:** Training on stale rollout data causes significant losses in training speed. Queue length estimates the amount of artificial staleness introduced; see subsection M.2 for experiment details.

(c) **Sample Reuse:** Reusing each sample of training data causes significant slowdowns. See subsection M.3 for experiment details.

Figure 5: **Batch Size and data quality in early training:** For each parameter, we ran multiple training runs varying only that parameter. These runs cover early training (approximately one week) at small scale (8x smaller than Rerun). On the left we plot TrueSkill over time for each run. On the right, we plot the “speedup” to reach fixed TrueSkill thresholds of 100, 125, 150, and 175 as a function of the parameter under study compared to the baseline (marked with ‘b’); see Equation 2. Higher speedup means that training was faster and more efficient. These four thresholds are chosen arbitrarily; a few are omitted when the uncertainties are too large (for example in Figure 5c fewer than half the experiments reach 175, so that speedup curve would not be informative).

portion of the game, and upload data to the experience buffer, while optimizers continually sample from whatever data is present in the experience buffer to optimize (Figure 2).

Early on in the project, we had rollout workers collect full episodes before sending it to the optimizers and downloading new parameters. This means that once the data finally enters the optimizers, it can be several hours old, corresponding to thousands of gradient steps. Gradients computed from these old parameters were often useless or destructive. In the final system rollout workers send data to optimizers after only 256 timesteps, but even so this can be a problem.

We found it useful to define a metric for this called *staleness*. If a sample was generated by parameter version N and we are now optimizing version M , then we define the *staleness* of that data to be $M - N$. In Figure 5b, we see that increasing staleness by ~ 8 versions causes significant slowdowns. Note that this level of staleness corresponds to a few minutes in a multi-month experiment. Our final system design targeted a staleness between 0 and 1 by sending game data every 30 seconds of gameplay and updating to fresh parameters approximately once a minute, making the loop faster than the time it takes the optimizers to process a single batch (32 PPO gradient steps). Because of the high impact of staleness, in future work it may be worth investigating whether optimization methods more robust to off-policy data could provide significant improvement in our asynchronous data collection regime.

Because optimizers sample from an experience buffer, the same piece of data can be re-used many times. If data is reused too often, it can lead to overfitting on the reused data[18]. To diagnose this, we defined a metric called the *sample reuse* of the experiment as the instantaneous ratio between the rate of optimizers consuming data and rollouts producing data. If optimizers are consuming samples twice as fast as rollouts are producing them, then on average each sample is being used twice and we say that the sample reuse is 2. In Figure 5c, we see that reusing the same data even 2-3 times can cause a factor of two slowdown, and reusing it 8 times may prevent the learning of a competent policy altogether. Our final system targets sample reuse ~ 1 in all our experiments.

These experiments on the early part of training indicate that high quality data matters even more than compute consumed; small degradations in data quality have severe effects on learning. Full details of the experiment setup can be found in Appendix M.

4.5 Long term credit assignment

Dota 2 has extremely long time dependencies. Where many reinforcement learning environment episodes last hundreds of steps ([4, 29–31]), games of Dota 2 can last for tens of thousands of time steps. Agents must execute plans that play out over many minutes, corresponding to thousands of timesteps. This makes our experiment a unique platform to test the ability of these algorithms to understand long-term credit assignment.

In Figure 6, we study the time horizon over which our agent discounts rewards, defined as

$$H = \frac{T}{1 - \gamma} \tag{3}$$

Here γ is the discount factor [17] and T is the real game time corresponding to each step (0.133 seconds). This measures the game time over which future rewards are integrated, and we use it as a proxy for the long-term credit assignment which the agent can perform.

In Figure 6, we see that resuming training a skilled agent using a longer horizon makes it perform better, up to the longest horizons we explored (6-12 minutes). This implies that our optimization was capable of accurately assigning credit over long time scales, and capable of learning policies and

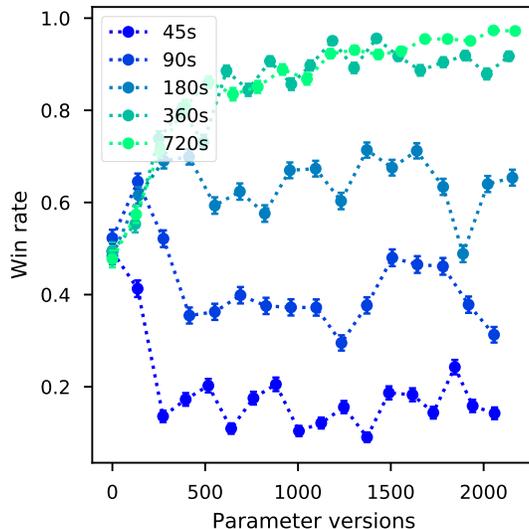


Figure 6: **Effect of horizon on agent performance.** We resume training from a trained agent using different horizons (we expect long-horizon planning to be present in highly-skilled agents, but not from-scratch agents). The base agent was trained with a horizon of 180 seconds ($\gamma = 0.9993$), and we include as a baseline continued training at horizon 180s. Increasing horizon increases win rate over the trained agent at the point training was resumed, with diminishing returns at high horizons.

actions which maximize rewards 6-12 minutes into the future. As the environments we attempt to solve grow in complexity, long-term planning and thinking will become more and more important for intelligent behavior.

5 Related Work

The OpenAI Five system builds upon several bodies of work combining deep reinforcement learning, large-scale optimization of deep learning models, and using self-play to explore environments and strategies.

Competitive games have long served as a testbed for learning. Early systems mastered Backgammon [1], Checkers [32], and Chess [2]. Self-play was shown to be a powerful algorithm for learning skills within high-dimensional continuous environments [33] and a method for automatically generating curricula [34]. Our use of self-play is similar in spirit to fictitious play [35], which has been successfully applied to poker [36] - in this work we learn a distribution over opponents and use the latest policy rather than an average policy.

Using a combination of imitation learning human games and self-play, Silver et al. demonstrated a master-level Go player [4]. Building upon this work, AlphaGoZero, AlphaZero, and ExIt discard imitation learning in favor of using Monte-Carlo Tree Search during training to obtain higher quality trajectories [12, 37, 38] and apply this to Go, Chess, Shogi, and Hex. Most recently, human-level play has been demonstrated in 3D first-person multi-player environments [30], professional-level play in the real-time strategy game StarCraft 2 using AlphaStar [7], and superhuman performance

in Poker [39].

AlphaStar is particularly relevant to this paper. In that effort, which ran concurrently to our own, researchers trained agents to play Starcraft 2, another complex game with real-time performance requirements, imperfect information, and long time horizons. The model for AlphaStar used a similar hand-designed architecture to embed observations and an autoregressive action decoder, with an LSTM core to handle partial observability. Both systems used actor critic reinforcement learning methods as part of the overall objective. OpenAI Five has certain sub-systems hard-coded (such as item buying), whereas AlphaStar handled similar decisions (e.g. building order) by conditioning (during training) on statistics derived from human replays. OpenAI Five trained using self play, while AlphaStar used a league consisting of multiple agents, where agents were trained to beat certain subsets of other agents. Finally, AlphaStar’s value network observed full information about the game state (including observations hidden from the policy); this method improved their training and exploring its application to Dota 2 is a promising direction for future work.

Deep reinforcement learning has been successfully applied to learning control policies from high dimensional input. In 2013, Mnih et al.[3] show that it is possible to combine a deep convolutional neural network with a Q-learning algorithm[40] and a novel experience replay approach to learn policies that can reach superhuman performance on the Atari ALE games. Following this work, a variety of efforts have pushed performance on the remaining Atari games[16], reduced the sample complexity, and introduced new challenges by focusing on intrinsic rewards [41–43].

As more computational resources have become available, a body of work has developed addressing the use of distributed systems in training. Larger batch sizes were found to accelerate training of image models[44–46]. Proximal Policy Optimization[14] and A3C [47] improve the ability to asynchronously collect rollout data. Recent work has demonstrated the benefit of distributed learning on a wide array of problems including single-player video games[48] and robotics[5].

The motivation for our surgery method is similar to prior work on *Net2Net* style function preserving transformations [23] which attempt to add model capacity without compromising performance, whereas our surgery technique was used in cases where the inputs, outputs, and recurrent layer size changed. Past methods have grown neural networks by incrementally training and freezing parts of the network [49], [50], [51]. Li & Hoiem [52] and Rusu *et al.* [53] use similar methods to use a trained model to quickly learn novel tasks. Distillation [54] and imitation learning [55, 56] offer an alternate approach to surgery for making model changes in response to a shifting environment. In concurrent work, OpenAI *et al.* [24] has reported success using behavioral cloning for similar purposes.

6 Conclusion

When successfully scaled up, modern reinforcement learning techniques can achieve superhuman performance in competitive esports games. The key ingredients are to expand the scale of compute used, by increasing the batch size and total training time. In order to extend the training time of a single run to ten months, we developed surgery techniques for continuing training across changes to the model and environment. While we focused on Dota 2, we hypothesize these results will apply more generally and these methods can solve any zero-sum two-team continuous environment which can be simulated in parallel across hundreds of thousands of instances. In the future, environments and tasks will continue to grow in complexity. Scaling will become even more important (for current methods) as the tasks become more challenging.

Acknowledgements

Myriad individuals contributed to this work from within OpenAI, within the Dota 2 community, and elsewhere. We extend our utmost gratitude to everyone who helped us along the way! We would like to especially recognize the following contributions:

- Technical discussions with numerous people within OpenAI including Bowen Baker, Paul Christiano, Danny Hernandez, Sam McCandlish, Alec Radford
- Review of early drafts by Bowen Baker, Danny Hernandez, Jacob Hilton, Quoc Le, Luke Metz, Matthias Plappert, Alec Radford, Oriol Vinyals
- Event Support from Larissa Schiavo, Diane Yoon, Loren Kwan
- Communication, writing, and outreach support from Ben Barry, Justin Wang, Shan Carter, Ashley Pilipiszyn, Jack Clark
- OpenAI infrastructure support from Eric Sigler
- Google Cloud Support (Solomon Boulos, JS Riehl, Florent de Goriainoff, Somnath Roy, Winston Lee, Andrew Sallaway, Danny Hammo, Jignesh Naik)
- Microsoft Azure Support (Jack Kabat, Jason Vallery, Niel Mackenzie, David Kalmin, Dina Frandsen)
- Dota 2 Support from Valve (special thanks to Chris Carollo)
- Dota 2 guides and builds from Tortedelini (Michael Cohen) and buyback saving strategy from Adam Michalik
- Dota 2 expertise and community advice from Blitz (William Lee)
- Dota 2 Casters: Blitz (William Lee), Capitalist (Austin Walsh), Purge (Kevin Godec), ODPixel (Owen Davies), Sheever (Jorien van der Heijden), Kyle Freedman
- Dota 2 World Champions (OG): ana (Anathan Pham), Topson (Topias Taavitsainen), Ceb (Sébastien Debs), JerAx (Jesse Vainikka), N0tail (Johan Sundstein)
- Dota 2 Professional Teams: Team Secret, Team Lithium, Alliance, SG E-sports
- Benchmark Players: Moonmeander (David Tan), Merlini (Ben Wu), Capitalist (Austin Walsh), Fogged (Ioannis Loucas), Blitz (William Lee)
- Playtesting: Alec Radford, Bowen Baker, Alex Botev, Pedja Marinkovic, Devin McGarry, Ryan Perron, Garrett Fisher, Jordan Beeli, Aaron Wasnich, David Park, Connor Mason, James Timothy Herron, Austin Hamilton, Kieran Wasylshyn, Jakob Roedel, William Rice, Joel Olazagasti, Samuel Anderson
- We thank the entire Dota 2 community for their support and enthusiasm. We especially profusely thank all 39,356 Dota 2 players from 225 countries who participated in OpenAI Five Arena and all the players who played against the 1v1 agent during the LAN event at The International 2017!

Author Contributions

This manuscript is the result of the work of the entire OpenAI Dota team. For each major area, we list the primary contributors in alphabetical order.

- Greg Brockman, Brooke Chan, Przemysław “Psyho” Dębiak, Christy Dennison, David Farhi, Scott Gray, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Szymon Sidor, Jie Tang, Filip Wolski, and Susan Zhang developed and trained OpenAI Five, including developing surgery, expanding tools for large-scale distributed RL, expanding the capabilities to the 5v5 game, and running benchmarks against humans including the OG match and OpenAI Arena.
- Christopher Berner, Greg Brockman, Vicki Cheung, Przemysław “Psyho” Dębiak, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Catherine Olsson, Jakub Pachocki, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, and Jie Tang developed the 1v1 training system, including the Dota 2 gym interface, building the first Dota agent, and initial exploration of batch size scaling.
- Brooke Chan, David Farhi, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Jie Tang, and Filip Wolski wrote this manuscript, including running Rerun and all of the ablation studies.
- Jakub Pachocki and Szymon Sidor set research direction throughout the project, including developing the first version of Rapid to demonstrate initial benefits of large scale computing in RL.
- Greg Brockman and Rafal Józefowicz kickstarted the team.

References

1. Tesauro, G. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation* **6**, 215–219 (1994).
2. Campbell, M., Hoane Jr., A. J. & Hsu, F.-h. Deep Blue. *Artif. Intell.* **134**, 57–83. ISSN: 0004-3702 (Jan. 2002).
3. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
4. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., *et al.* Mastering the game of Go with deep neural networks and tree search. *nature* **529**, 484 (2016).
5. OpenAI. *Learning Dexterity* <https://openai.com/blog/learning-dexterity/>. [Online; accessed 28-May-2019]. 2018.
6. Paulus, R., Xiong, C. & Socher, R. *A Deep Reinforced Model for Abstractive Summarization* 2017. arXiv: 1705.04304 [cs.CL].
7. Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., *et al.* Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 1–5 (2019).
8. Guss, W. H., Codel, C., Hofmann, K., Houghton, B., Kuno, N., Milani, S., Mohanty, S. P., Liebana, D. P., Salakhutdinov, R., Topin, N., Veloso, M. & Wang, P. The MineRL Competition on Sample Efficient Reinforcement Learning using Human Priors. *CoRR* **abs/1904.10079**. arXiv: 1904.10079. <http://arxiv.org/abs/1904.10079> (2019).
9. Wikipedia contributors. *Dota 2 — Wikipedia, The Free Encyclopedia* https://en.wikipedia.org/w/index.php?title=Dota_2&oldid=913733447. [Online; accessed 9-September-2019]. 2019.
10. Wikipedia contributors. *The International 2018 — Wikipedia, The Free Encyclopedia* https://en.wikipedia.org/w/index.php?title=The_International_2018&oldid=912865272. [Online; accessed 9-September-2019]. 2019.
11. Allis, L. V. *Searching for solutions in games and artificial intelligence* in (1994).
12. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., *et al.* A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**, 1140–1144 (2018).
13. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with LSTM (1999).
14. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
15. Konda, V. R. & Tsitsiklis, J. N. *Actor-critic algorithms* in *Advances in neural information processing systems* (2000), 1008–1014.
16. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. & Kavukcuoglu, K. *Asynchronous methods for deep reinforcement learning* in *International conference on machine learning* (2016), 1928–1937.

17. Schulman, J., Moritz, P., Levine, S., Jordan, M. I. & Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *CoRR* **abs/1506.02438** (2016).
18. Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., van Hasselt, H. & Silver, D. Distributed Prioritized Experience Replay. *CoRR* **abs/1803.00933**. arXiv: 1803.00933. <http://arxiv.org/abs/1803.00933> (2018).
19. NVIDIA. *NVIDIA Collective Communications Library (NCCL)* <https://developer.nvidia.com/nccl>. [Online; accessed 9-September-2019]. 2019.
20. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
21. Williams, R. J. & Peng, J. An Efficient Gradient-Based Algorithm for On-Line Training of Recurrent Network Trajectories. *Neural Computation* **2**, 490–501 (1990).
22. Gray, S., Radford, A. & Kingma, D. P. *GPU Kernels for Block-Sparse Weights* 2017.
23. Chen, T., Goodfellow, I. J. & Shlens, J. *Net2Net: Accelerating Learning via Knowledge Transfer in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (2016). <http://arxiv.org/abs/1511.05641>.
24. OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W. & Zhang, L. *Solving Rubik’s Cube with a Robot Hand* 2019. arXiv: 1910.07113 [cs.LG].
25. Dalvi, N., Domingos, P., Mausam, Sanghai, S. & Verma, D. *Adversarial Classification in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, Seattle, WA, USA, 2004)*, 99–108. ISBN: 1-58113-888-1. doi:10.1145/1014052.1014066. <http://doi.acm.org/10.1145/1014052.1014066>.
26. Jain, A., Bansal, R., Kumar, A. & Singh, K. A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International journal of applied and basic medical research* **5**, 125–127 (2015).
27. Herbrich, R., Minka, T. & Graepel, T. *TrueSkill: a Bayesian skill rating system in Advances in neural information processing systems* (2007), 569–576.
28. McCandlish, S., Kaplan, J., Amodei, D. & Team, O. D. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162* (2018).
29. Cobbe, K., Klimov, O., Hesse, C., Kim, T. & Schulman, J. Quantifying Generalization in Reinforcement Learning. *CoRR* **abs/1812.02341**. arXiv: 1812.02341. <http://arxiv.org/abs/1812.02341> (2018).
30. Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., *et al.* Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv preprint arXiv:1807.01281* (2018).
31. Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M. & Bowling, M. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* **356**, 508–513 (2017).

32. Schaeffer, J., Culberson, J., Treloar, N., Knight, B., Lu, P. & Szafron, D. A world championship caliber checkers program. *Artificial Intelligence* **53**, 273–289. ISSN: 0004-3702 (1992).
33. Bansal, T., Pachocki, J., Sidor, S., Sutskever, I. & Mordatch, I. Emergent complexity via multi-agent competition. *arXiv preprint arXiv:1710.03748* (2017).
34. Sukhbaatar, S., Lin, Z., Kostrikov, I., Synnaeve, G., Szlam, A. & Fergus, R. Intrinsic motivation and automatic curricula via asymmetric self-play. *arXiv preprint arXiv:1703.05407* (2017).
35. Brown, G. W. in *Activity Analysis of Production and Allocation* (ed Koopmans, T. C.) (Wiley, New York, 1951).
36. Heinrich, J. & Silver, D. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. *CoRR* **abs/1603.01121**. arXiv: 1603.01121 (2016).
37. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., *et al.* Mastering the game of go without human knowledge. *Nature* **550**, 354 (2017).
38. Anthony, T., Tian, Z. & Barber, D. *Thinking fast and slow with deep learning and tree search* in *Advances in Neural Information Processing Systems* (2017), 5360–5370.
39. Brown, N. & Sandholm, T. Superhuman AI for multiplayer poker. *Science*, eaay2400 (2019).
40. Watkins, C. J. & Dayan, P. Q-learning. *Machine learning* **8**, 279–292 (1992).
41. Kulkarni, T. D., Narasimhan, K., Saeedi, A. & Tenenbaum, J. *Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation* in *Advances in neural information processing systems* (2016), 3675–3683.
42. Burda, Y., Edwards, H., Storkey, A. & Klimov, O. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* (2018).
43. Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O. & Clune, J. Montezuma’s revenge solved by go-explore, a new algorithm for hard-exploration problems (sets records on pitfall too). *Uber Engineering Blog, Nov* (2018).
44. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y. & He, K. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour (June 2017).
45. You, Y., Gitman, I. & Ginsburg, B. Scaling SGD Batch Size to 32K for ImageNet Training (Aug. 2017).
46. You, Y., Zhang, Z., Hsieh, C.-J., Demmel, J. & Keutzer, K. *ImageNet Training in Minutes* in *Proceedings of the 47th International Conference on Parallel Processing* (ACM, Eugene, OR, USA, 2018), 1:1–1:10. ISBN: 978-1-4503-6510-9. doi:10.1145/3225058.3225069. <http://doi.acm.org/10.1145/3225058.3225069>.
47. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. & Kavukcuoglu, K. *Asynchronous Methods for Deep Reinforcement Learning* in *Proceedings of The 33rd International Conference on Machine Learning* (eds Balcan, M. F. & Weinberger, K. Q.) **48** (PMLR, New York, New York, USA, June 2016), 1928–1937. <http://proceedings.mlr.press/v48/mnih16.html>.
48. Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., *et al.* Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561* (2018).

49. Fahlman, S. E. & Lebiere, C. in (ed Touretzky, D. S.) 524–532 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990). ISBN: 1-55860-100-7. <http://dl.acm.org/citation.cfm?id=109230.107380>.
50. Wang, Y., Ramanan, D. & Hebert, M. *Growing a Brain: Fine-Tuning by Increasing Model Capacity* in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017* (2017), 3029–3038. doi:10.1109/CVPR.2017.323. <https://doi.org/10.1109/CVPR.2017.323>.
51. Czarnecki, W. M., Jayakumar, S. M., Jaderberg, M., Hasenclever, L., Teh, Y. W., Osindero, S., Heess, N. & Pascanu, R. Mix&Match - Agent Curricula for Reinforcement Learning. *CoRR* **abs/1806.01780**. arXiv: 1806.01780. <http://arxiv.org/abs/1806.01780> (2018).
52. Li, Z. & Hoiem, D. Learning without Forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**, 2935–2947. ISSN: 0162-8828 (Dec. 2018).
53. Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R. & Hadsell, R. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671* (2016).
54. Hinton, G., Vinyals, O. & Dean, J. *Distilling the Knowledge in a Neural Network* in *NIPS Deep Learning and Representation Learning Workshop* (2015). <http://arxiv.org/abs/1503.02531>.
55. Ross, S., Gordon, G. & Bagnell, D. *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning* in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (eds Gordon, G., Dunson, D. & Dudík, M.) **15** (PMLR, Fort Lauderdale, FL, USA, Apr. 2011), 627–635. <http://proceedings.mlr.press/v15/ross11a.html>.
56. Levine, S. & Koltun, V. *Guided Policy Search* in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (JMLR.org, Atlanta, GA, USA, 2013), III-1–III-9. <http://dl.acm.org/citation.cfm?id=3042817.3042937>.
57. OpenAI. *AI and Compute* <https://openai.com/blog/ai-and-compute/>. [Online; accessed 9-Sept-2019]. 2018.
58. Ng, A. Y., Harada, D. & Russell, S. *Policy invariance under reward transformations: Theory and application to reward shaping* in *In Proceedings of the Sixteenth International Conference on Machine Learning* (Morgan Kaufmann, 1999), 278–287.
59. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. & Zaremba, W. OpenAI Gym. *CoRR* **abs/1606.01540**. arXiv: 1606.01540. <http://arxiv.org/abs/1606.01540> (2016).
60. Balduzzi, D., Garnelo, M., Bachrach, Y., Czarnecki, W. M., Pérolat, J., Jaderberg, M. & Graepel, T. Open-ended Learning in Symmetric Zero-sum Games. *CoRR* **abs/1901.08106**. arXiv: 1901.08106. <http://arxiv.org/abs/1901.08106> (2019).
61. Williams, R. J. & Peng, J. Function optimization using connectionist reinforcement learning algorithms. *Connection Science* **3**, 241–268 (1991).

Appendix

Table of Contents

A	Compute Usage	25
B	Surgery	25
C	Hyperparameters	29
D	Evaluating agents' understanding	30
D.1	Understanding OpenAI Five Finals	33
D.2	Hero selection	35
E	Observation Space	37
F	Action Space	39
F.1	Scripted Actions	42
G	Reward Weights	43
H	Neural Network Architecture	45
I	Human Games	49
J	TrueSkill: Evaluating a Dota 2 Agent Automatically	49
K	Dota 2 Gym Environment	49
K.1	Data flow between the training environment and Dota 2	49
L	Reaction time	51
M	Scale and Data Quality Ablation Details	52
M.1	Batch Size	53
M.2	Sample Quality — Staleness	54
M.3	Sample Quality — Sampling and Sample Reuse	56
N	Self-play	59
O	Exploration	59
O.1	Loss function	59

O.2 Environment Randomization	62
P Hero Pool Size	63
Q Bloopers	65
Q.1 Manually Tuned Hyperparameters	65
Q.2 Zero Team Spirit Embedding	65
Q.3 Learning path dependency	66

A Compute Usage

We estimate the optimization compute usage as follows: We break the experiment in segments between each major surgery or batch size change. For each of those, we calculate the number of gradient descent steps taken (number of iterations \times 32). We estimate the compute per step per GPU using TensorFlow’s `tf.profiler.total_float_ops`, then multiply together:

$$\text{total compute} = \sum_{\text{segment}} 32 \times (\text{iteration}_{\text{end}} - \text{iteration}_{\text{start}}) \times \quad (4)$$

$$(\# \text{ gpus}) \times (\text{compute per step per gpu}) \quad (5)$$

Our uncertainty on this estimate comes primarily from ambiguities about what computation “counts.” For example the tensorflow metrics include all ops in the graph including metric logging, nan-checking, etc. It also includes the prediction of auxiliary heads such as win probability, which are not necessary for gameplay or training. It does not count non-GPU compute on the optimizer machines such as exporting parameter versions to the rollouts. We estimate these and other ambiguities to be around 5%. In addition, for OpenAI Five (although not for Rerun) we use a simplified history of the experiment, rather than keeping track of every change and every time something crashed and needed to be restarted; we estimate this does not add more than 5% error. We combine these rough error estimates into a (very crude) net ambiguity estimate of 5-10%.

This computation concludes that OpenAI Five used 770 ± 50 PFlops/s-days of total optimization compute on GPUs at the time of playing the world champions (April 13, 2019), and 820 ± 50 total optimization compute when it was finally turned off on April 22nd, 2019. Rerun, on the other hand, used 150 ± 5 PFlops/s-days between May 18th and July 12th, 2019.

We adopted the methodology from [57] to facilitate comparisons. This has several important caveats. First, the above computation only considers compute used for optimization. In fact this is a relatively small portion of the total compute budget for the training run. In addition to the GPU machines doing optimization (roughly 30% of the cost by dollars spent) there are approximately the same number of GPUs running forward passes for the rollout workers (30%), as well as the actual rollouts CPUs running the selfplay games (30%) and the overhead of controllers, TrueSkill evaluators, CPUs on the GPU machines, etc (10%).

Second, with any research project one needs to run many small studies, ablations, false starts, etc. One also inevitably wastes some computing resources due to imperfect utilization. Traditionally the AI community has not counted these towards the compute used by the project, as it is much easier to count only the resources used by the final training run. However, with our advent of surgery, the line becomes much fuzzier. After 5 months of training on an older environment, we *could* have chosen to start from scratch in the new environment, or performed surgery to keep the old model. Either way, the same total amount of compute gets used; but the above calculation ignores all the compute used up until the last time we chose to restart. For these reasons the compute number for OpenAI Five should be taken with a large grain of salt, but this caveat does not apply to Rerun, which was trained without surgery.

B Surgery

As discussed in 3.3, we designed “surgery” tools for continuing to train a single set of paramters across changes to the environment, model architecture, observation space, and action space. The

Date	Iteration	# params	Change
6/30/2018	1	43,436,520	Experiment started
8/17/2018	81,821	43,559,322	Dota 2 version 7.19 adds new items, abilities, etc.
8/18/2018	84,432	43,805,274	Change environment to single courier; remove “cheating” observations
8/26/2018	91,471	156,737,674	Double LSTM size
9/27/2018	123,821	156,809,485	Support for more heroes
10/3/2018	130,921	156,809,501	Obs: Roshan spawn timing
10/12/2018	140,402	156,811,805	Item: Bottle
10/19/2018	144,121	156,286,925	Obs: Stock counts; Obs: Remove some obsolete obs
10/24/2018	150,111	156,286,867	Obs: Neutral creep & rune spawn timers
11/7/2018	161,482	156,221,309	Obs: Item swap cooldown; Obs: Remove some obsolete obs
11/28/2018	185,749	156,221,669	Item: Divine rapier; Obs: Improve observation of stale enemy heroes
12/10/2018	193,701	157,378,165	Obs: Modifiers on nonhero units.
12/14/2018	196,800	157,650,795	Action: Consumables on allies; Obs: Line of sight information; Obs: next item this hero will purchase; Action: buyback
12/20/2018	203,241	157,679,655	Dota 2 version 7.20 adds new items, new item slot, changes map, etc; Obs: number of empty inventory slots
1/23/2019	211,191	158,495,991	Obs: Improve observations of area of effects; Obs: improve observation of modifiers’ duration; Obs: Improve observations about item Power Treads.
4/5/2019	220,076	158,502,815	Dota 2 version 7.21 adds new items, abilities, etc.

Table 1: All successful surgeries and major environment changes performed during the training of OpenAI Five. This table does not include surgeries which were ultimately reverted due to training failures, nor minor environment changes (such as improvements to partial reward weights or scripted logic). “Obs” indicates that a new observation was added as an input to the model or an existing one was changed. “Action” indicates that a new game action was made available, along with appropriate observations about the state of that action. “Item” indicates that a new item was introduced, including observation of the item and the action to use the item. The Dota 2 version updates (7.19, 7.20 and 7.21) include many new items, actions, and observations.

goal in each case is to resume training after the change without the agent losing any skill from the change. Table 1 lists the major surgeries we performed in the lifetime of the OpenAI Five experiment.

For changes which add parameters, one of the key questions to ask is how to initialize the new parameters. If we initialize the parameters randomly and continue optimization, then noise will flow into other parts of the model, causing the model to play badly and causing large gradients which destroy the learned behaviors.

In the rest of this appendix we provide details of the tools we used to continue training across each type of change. In general we had a high-skill model π_θ trained to act in one environment, and due to a change to the problem design we need to begin training a newly-shaped model $\hat{\pi}_{\hat{\theta}}$ in a new environment. Ultimately the goal is for the TrueSkill of agent $\hat{\pi}_{\hat{\theta}}$ to match that of π_θ .

Changing the architecture In the most straightforward situation, the observation space, action space, and environment do not change. In this case, per Equation 1, we can insist that the new policy $\hat{\pi}_{\hat{\theta}}$ implement exactly the same mathematical function from observations to actions as the old policy.

A simple example here would be adding more units to an internal fully-connected layer of the model. Suppose that before the change, some part of the interior of the model contained an input vector x (dimension d_x), which is transformed to an activation vector $y = W_1x + B_1$ (dimension d_y), which is then consumed by another fully-connected layer $z = W_2y + B_2$ (dimension d_z). We desire to increase the dimension of y from d_y to \hat{d}_y . This causes the shapes of three parameter arrays to change: W_1 (from $[d_x, d_y]$ to $[d_x, \hat{d}_y]$), B_1 (from $[d_y]$ to $[\hat{d}_y]$), and W_2 (from $[d_y, d_z]$ to $[\hat{d}_y, d_z]$).

In this case we initialize the new variables in the first layer as:

$$\hat{W}_1 = \begin{bmatrix} W_1 \\ R() \end{bmatrix} \quad \hat{B}_1 = \begin{bmatrix} B_1 \\ R() \end{bmatrix} \quad \hat{W}_2 = [W_2 \quad 0] \quad (6)$$

Where $R()$ indicates a random initialization. The initializations of \hat{W}_1 and \hat{B}_1 ensure that the first d_y dimensions of activations \hat{y} will be the same data as the old activations y , and the remained will be randomized. The randomization ensures that symmetry is broken among the new dimensions. The initialization of \hat{W}_2 , on the other hand, ensures that the next layer will ignore the new random activations, and the next layer’s activations will be the same as in the old model; $\hat{z} = z$. The weights which are initialized to zero will move away from zero due to the gradients, if the corresponding new dimensions in y are useful to the downstream function.

Initializing neural network weights to zero is a dangerous business, because it can introduce undesired symmetries between the indices of the output vector. However we found that in most cases of interest, this was easy to avoid by only zero-ing the minimal set of weights. In the example above, the symmetry is broken by the randomization of \hat{W}_1 and \hat{B}_1 .

A more advanced version of this surgery was required when we wanted to increase the model capacity dramatically, by increasing the hidden dimension of our LSTM from 2048 units to 4096 units. Because the LSTM state is recurrent, there was no way to achieve the separation present in Equation 6; if we randomize the new weights they will impact performance, but if we set them to zero then the new hidden dimensions will be symmetric and gradient updates will never differentiate them. In practice we set the new weights to random small values — rather than randomize new weight values on the same order of magnitude as the existing weights, we randomized new weights

significantly smaller. The scale of randomization was set empirically by choosing the highest scale which did not noticeably decrease the agent’s TrueSkill.

Changing the Observation Space Most of our surgeries caused the observation space changes, for example when we added 3 new float observations encoding the time until neutral creeps, bounties, and runes would spawn. In these cases it is impossible to insist that the new policy implement the same function from observation space to action space, as the input domain has changed. However, in some sense the input domain has *not* changed; the game state is still the same. In reality our system is not only a function $\pi : o \rightarrow a$; before the policy sees the observation arrays, an “encoder” function E has turned a game state s into an input array o :

$$(\text{Game State Protobuf } s) \xrightarrow{E} (\text{Observation Arrays } o) \xrightarrow{\pi} (\text{Action } a) \tag{7}$$

By adding new observations we are enhancing the encoder function E , making it take the same game state and simply output richer arrays for the model to consume. Thus in this case while we cannot ensure that $\hat{\pi}_{\hat{\theta}} = \pi_{\theta}$, we can ensure the functions are identical if we go one step back:

$$\forall s \quad \hat{\pi}_{\hat{\theta}}(\hat{E}(s)) = \pi_{\theta}(E(s)) \tag{8}$$

When the change is simply additive, this can then be enforced as in the previous section. Suppose the new observations extend a vector x from dimension d_x to dimension \hat{d}_x , and the input vector x is consumed by a weight matrix W via $y = Wx$ (and y is then processed by the rest of the model downstream). Then we initialize the new weights \hat{W} as:

$$\hat{W} = [W \quad 0] \tag{9}$$

As before, this ensures that the rest of the model is unchanged, as the output is unchanged ($\hat{y} = y$). The weights which are initialized to zero will move away from zero due to the gradients, if the corresponding observations are found to be useful.

Changing the Environment or Action Space The second broad class of changes are those which change the environment itself, either by making new actions available to the policy (e.g. when we replaced scripted logic for the Buyback action with model-controlled logic) or by simply changing the Dota 2 rules (for example when we moved to Dota 2 version 7.21, or when we added new items). For some of these changes, such as upgrading Dota 2 version, we found simply making the change on the rollout workers to be relatively stable; the old policy played well enough in the new environment that it was able to smoothly adapt.

Even so, whenever possible, we attempted to “anneal” in these new features, starting with 0% of rollout games played with the new environment or actions, and slowly ramping up to 100%. This prevents a common problem where a change in one part of the agent’s behavior could force unnecessary relearning large portions of the strategy. For example, when we attempted to give the model control of the Buyback action without annealing, the model-based control of the action was (at first) worse than the scripted version had been, causing the agent to adapt its overall strategies to games where allies and enemies alike often misuse this action. This would cause the agent to significantly drop in overall skill; while it would likely eventually recover, it may require “repeating” the investment of a large amount of compute. By annealing the new action in gradually, we ensure that the model never loses overall skill due to a sudden change of one part of the environment;

when we observe the model losing TrueSkill during the annealing process, we revert and attempt the anneal at a slower rate. This annealing process makes sense even if the environment is becoming fundamentally “harder” because our agent’s skill is measured through winrates against other models; the opponent also has to play in the new environment.

Removing Model Parts Requiring exact policy equivalence after the surgery outlaws many types of surgery. For example, most surgeries which remove parameters are not possible in this framework. For this reason our model continued to observe some “deprecated” observations, which were simply always set to constants. Further work such as [24] has already begun to explore alternate methods of surgery which avoid this constraint.

Smooth Training Restart The gradient moments stored by the Adam optimizer present a nuisance when restarting training with new parameter shape. To ensure that the moments have enough time to properly adjust, we use a learning rate of 0 for the first several hours of training after surgery. This also ensures that the distribution of rollout games has entered steady state by the time we begin training in earnest.

One additional nuisance when changing the shape of the model is the entire history of parameters which are stored (in the past opponent manager, see Appendix N), and used as opponents in rollouts. Because the rollout GPUs will be running the newest code, all of these past versions must be updated in the same way as the current version to ensure compatibility. If the surgery operation fails to exactly preserve the policy function, these frozen past agents will forever play worse, reducing the quality of the opponent pool. Therefore it is crucial to ensure agent behavior is unchanged after surgery.

Benefits of Surgery These surgeries primarily permitted us to have a tighter iteration loop for these features. When we added a new game feature which we expect to only matter at high skill, it would simply be impossible to test and iterate on it by training from scratch. Using surgery from the current OpenAI Five, we could have a more feasible process, which allowed us to safely include many minor features and improvements that otherwise would have been impossible to verify, such as adding long-tail items (Bottle, Rapier), minor improvements to the observation space (stock counts, modifiers on nonheroes), and others.

C Hyperparameters

The optimization algorithm has several important hyperparameters that have different settings throughout the training process. Over the course of training of OpenAI Five, these hyperparameters were modified by looking for improvement plateaus. Because of compute limitations which prevented us from testing hyperparameter changes in separate experiments, OpenAI Five’s long-running training process included numerous experimental hyperparameter changes. Some of these worked well and were kept, others were reverted as our understanding developed over the course of the 10-month period. As it is impossible for us to scan over any of these hyperparameters when our experiment is so large, we make no claim that the hyperparams used are optimal.

When we ran Rerun we simplified the hyperparameter schedule based on the lessons we had learned. In the end we made changes to only four key hyperparameters:

Iteration	0	15k	23k	43k	54k
Time (days)	0	13	20	33	42
TrueSkill	0	210	232	245	258
Team Spirit	0.3	0.8			
GAE Horizon	180 secs		360 secs		
Entropy coefficient	0.01			0.001	
Learning Rate	5e-5				5e-6

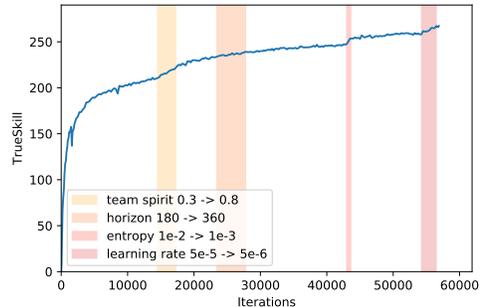


Figure 7: Hyperparameter changes during Rerun. Changes are displayed in table-form on the left, and called out in the trueskill vs iterations graph of the training run on the right. Each hyperparameter change was applied gradually over the course of 1-2 days, corresponding to several thousand iterations (the reported time in the table is the start of the change). Our pre-planned schedule included further changes to bring the experiment into line with OpenAI Five’s final hyperparameters (Horizon to 840 sec, team spirit to 1.0, and learning rate to 1e-6), but Rerun reached OpenAI Five’s skill level before we reached those hyperparameters.

- Learning Rate
- Entropy penalty coefficient (see Appendix O)
- Team Spirit (see Appendix G)
- GAE time horizon (see Equation 3)

This schedule is far from optimized as it was used in only our second iteration of this large experiment. In future work it could likely be significantly improved.

There are many other hyperparams that were not changed during the final Rerun experiment. Their values are listed in Table 2. Some of these were changed in the original OpenAI Five out of necessity (e.g. batch size changed many times as more or less compute resources became available, or SampleReuse changed as the relative speeds of different machine pools fluctuated), and others were changed experimentally in the original OpenAI Five run but were ultimately not important as evidenced by Rerun working without those changes (e.g. increasing the time horizon from 360 seconds to 840 seconds).

D Evaluating agents’ understanding

It is often difficult to infer the intentions of an RL agent. Some actions are obviously useful — hitting an enemy that is low on health, or freezing them as they’re trying to escape — but many other decisions can be less obvious. This is tightly coupled with questions on intentionality: does our agent plan on attacking the tower, or does it opportunistically deal the most damage possible in next few seconds?

To assess this, we attempt to predict future state of various features of the game from agent’s LSTM state:

- **Win probability:** Binary label of either 0 or 1 at the end of the game.

Param	Rerun	OpenAI Five	Baseline
Frameskip ^e	4	4	4
LSTM Unroll length ^e	16	16	16
Samples Per Segment ^e	16	16	16
Number of optimizer GPUs	512	480↔1,536	64
Batch Size/optimizer GPU (samples)	120	120↔128	120
Total Batch Size (samples) ^a	61,440	61,440↔196,608	7,680
Total Batch Size (timesteps) ^a	983,040	983,040↔3,145,728	122,880
Number of rollout GPUs	512	500↔1,440	64
Number of rollout CPUs	51,200	80,000↔172,800	6,400
Steps per Iteration	32	32	32
LSTM Size	4096	2048 → 4096	4096
Sample Reuse	1.0 ↔ 1.1	0.8↔2.7	1.0↔1.1
Team Spirit	0.3 → 0.8	0.3 → 1.0	0.3
GAE Horizon	180 secs → 360 secs	60 secs → 840 secs	180 secs
GAE λ	0.95	0.95	0.95
PPO clipping	0.2	0.2	0.2
Value loss weight ^c	1.0	0.25 ↔ 1.0	1.0
Entropy coefficient	0.01 → 0.001	0.01 → 0.001	0.01
Learning rate	5e-5 → 5e-6	5e-5 ↔ 1e-6	5e-5
Adam β_1	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.999
Past opponents ^b	20%	20%	20%
Past Opponents Learning Rate ^d	0.01	0.01	0.01

^a Batch size can be measured in samples (each an unrolled LSTM of 16 frames) or in individual timesteps.

^b Fraction of games played against past opponents (as opposed to self-play).

^c We normalize rewards using a running estimate of the standard deviation, and the value loss weight is applied post-normalization.

^d See Appendix N.

^e See Figure 8 for definitions of the various timescale subdivisions of a rollout episode.

Table 2: Hyperparameters: The OpenAI Five and Rerun columns indicate what was done for those individual experiments. For those which were modified during training, $x \rightarrow y$ indicates a smooth monotonic transition (usually a linear change over one to three days), and $x \leftrightarrow y$ indicates a less controlled variation due to either ongoing experimentation or distributed systems fluctuations. The “Baseline” indicates the default values for all the experiments in Appendix M (each individual experiment used these hyperparameters other than any it explicitly studied, for example in subsection M.1 the batch size was changed from the baseline in each training run, but all the other hyperparameters were from this table).

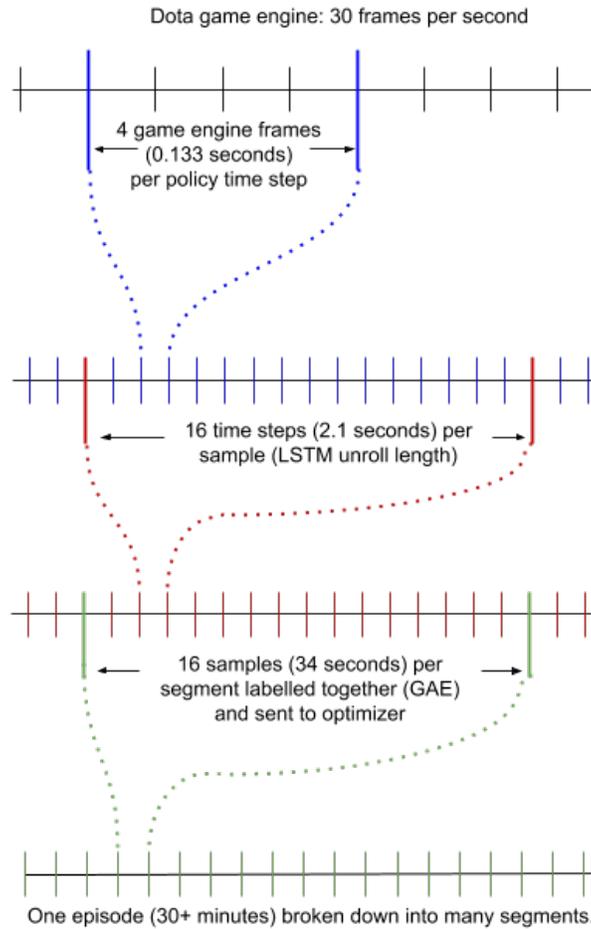


Figure 8: **Timescales and Staleness:** The breakdown of a rollout game. Rather than collect an entire game before sending it to the optimizers, rollout machines send data in shorter segments. The segment is further subdivided into samples of 16 policy actions which are optimized together using truncated BPTT. Each policy action bundles together four game engine frames.

- **Net worth rank:** Which rank among the team (1-5) in terms of total resources collected will this hero be at the end of the game? This prediction is used by scripted item-buying logic to decide which agents buy items shared by the team such as wards. In human play (which the scripted logic is based on) this task is traditionally performed by heroes who will have the lowest net worth at the end of the game.
- **Team objectives / enemy buildings:** whether this hero will help the team destroy a given enemy building in the near future.

We added small networks of fully-connected layers that transform LSTM output into predictions of these values. For historical reasons, win probability passes gradients to the main LSTM and rest of the agent with a very small weight; the other auxiliary predictions use Tensorflow’s `stop_gradient` method to train on their own.

One difficulty in training these predictors is that we train our agent on 30-second segments of the game (see Figure 8), and any given 30-second snippet may not contain the ground truth (e.g. for win probability and networth position, we only have ground truth on the very last segment of the game). We address this by training these heads in a similar fashion to how we train value functions. If a segment contains the ground truth label, we use the ground truth label for all time steps in that segment; if not, we use the model’s prediction at the end of the segment as the label. For win probability, for example, more precisely the label y for a segment from time t_1 to t_2 is given by:

$$y = \begin{cases} 1 & \text{last segment of the game, we win} \\ 0 & \text{last segment of the game, we lose} \\ \hat{y}(t_2) & \text{else} \end{cases} \quad (10)$$

Where $\hat{y}(t_2)$ is the model’s predicted win probability at the end of the segment. Although this requires information to travel backward through the game, we find it trains these heads to a degree of calibration and accuracy.

For the team objectives, we are additionally interested in whether the event will happen soon. For these we apply an additional discount factor with horizon of 2 minutes. This means that the enemy building predictions are not calibrated probabilities, but rather probabilities discounted by the expected time to the event.

D.1 Understanding OpenAI Five Finals

We used these supervised predictions to look closer at the game 1 from OpenAI Five Finals.

In Figure 9 we explore the progression of win probability predictions over the course of training Rerun, illustrating the evolution of understanding. Version 5,000 of the agent (early in the training process and low performance) already has a sense of what situations in the game may lead to eventual win. The prediction continues to get better and better as training proceeds. This matches human performance at this task, where even spectators with relatively little gameplay experience can estimate who is ahead based on simple heuristics, but with more gameplay practice human experts can estimate the winner more and more accurately.

On the winrate graph two dramatic game events are marked, at roughly the 5 and 18 minute point. One of them illustrates OpenAI Five’s win probability drop, due to an unexpected loss of 3 heroes in close succession. The other shows how the game turns from good to great as a key enemy hero is killed.

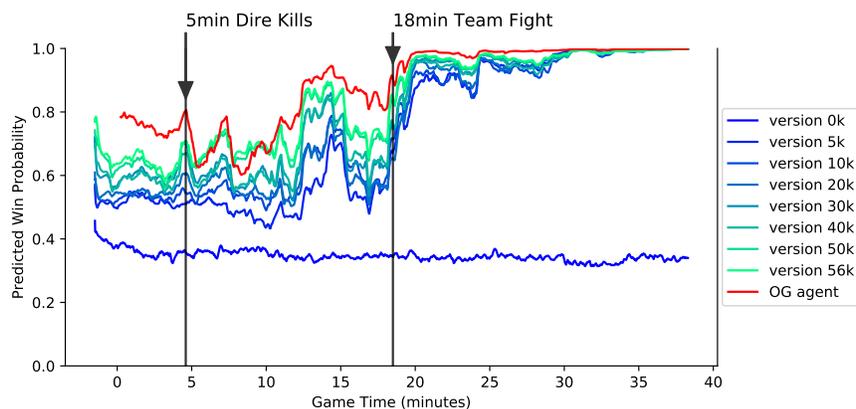


Figure 9: **Win Probability prediction of game 1 of OpenAI Five Finals** In red we show the (OpenAI Five) agent’s win probability prediction over the course of the game (which can be viewed by downloading the replay from <https://openai.com/blog/how-to-train-your-openai-five/>). Marked are two significant events that significantly affected win probability prediction. At roughly 5 minutes in the human team killed several of OpenAI Five’s heroes, making it doubt its lead. At roughly 18 minutes in, OpenAI Five team killed three human heroes in a row, regrouped all their heroes at the mid lane, and marched on declaring 95% probability of victory. Versions 0-56k are progressive versions of Rerun agent predicting win probabilities by replaying the same game; as we can see, prediction converges to that of the bot that actually played the game (original OpenAI Five), despite training over self-play games from separate training runs.

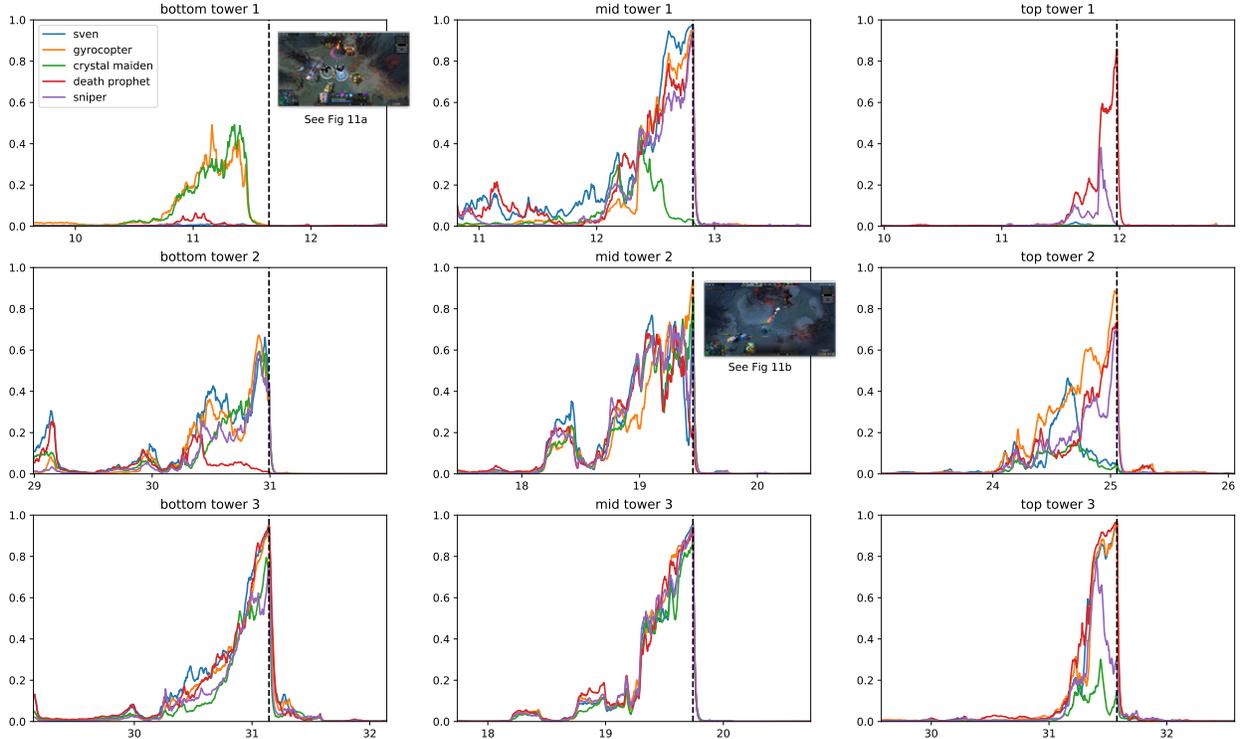


Figure 10: Continuous prediction of destroying enemy buildings by OpenAI Five in Finals game 1. Predictions by different heroes differ as they specifically predict whether they will participate in bringing given building down. Predictions should not be read as calibrated probabilities, because they are trained with a discount factor. See Figure 11a and Figure 11b for descriptions of the events corresponding to two of these buildings.

We also looked at heroes participation in destroying objectives. In Figure 10 we can see different heroes' predictions for each of the objectives in game 1 of OpenAI Five Finals. In several cases all heroes predict they will participate in the attack (and they do). In few cases one or two heroes are left out, and indeed by watching the game replay we see that those heroes are busy in the different part of the map during that time. In Figure 11 we illustrate these predictions with more details for two of the events.

D.2 Hero selection

In the normal game of Dota 2, two teams at the beginning of the game go through the process of selecting heroes. This is a very important step for future strategy, as heroes have different skill sets and special abilities. OpenAI Five, however, is trained purely on learning to play the best game of Dota 2 possible given randomly selected heroes.

Although we could likely train a separate drafting agent to play the draft phase, we do not need to; instead we can use the win probability predictor. Because the main varying observation that agents see at the start of the game is which heroes are on each team, the win probability at the start of the game estimates the strength of a given matchup. Because there are only 4,900,896 combinations of two 5-hero teams from the pool of 17 heroes, we can precompute agent's predicted



(a)



(b)

Figure 11: Screenshots right before two of the dire tower falls in the OpenAI Five Finals game 1. In 11a, Gyrocopter and Crystal Maiden attack the bottom tower 1 (upper left in Figure 10) and plan perhaps to kill it (their predictions go up). But they are chased away by the incoming dire (human) heroes, and their plan changes (the prediction that they will participate in the tower kill falls back to zero). Radiant creeps kill the tower half a minute later. In 11b, all radiant heroes attack mid tower 2 (center in Figure 10). However just before it falls, few dire heroes show up trying to save it, and most radiant heroes end up chasing them a fair distance away from the building. The prediction for those heroes to participate in the tower kill drops accordingly.

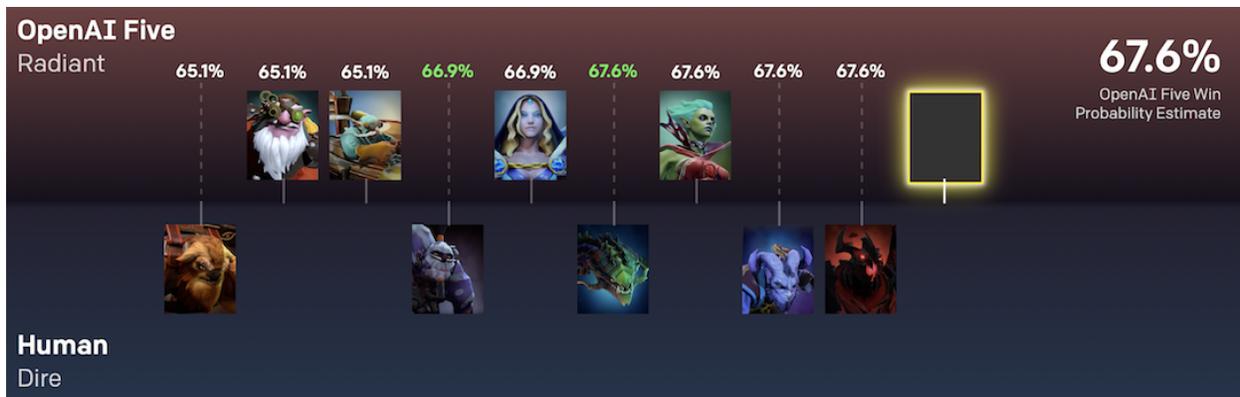


Figure 12: When drafting heroes, our drafting program would pick the one that maximizes worst-case scenario of opponent hero selection (minimax algorithm). In this example (from OpenAI Finals game 1), OpenAI Five deems the humans’ first pick suboptimal, immediately updating its expected win probability of 52.8% to 65.1%. The drafter then makes two choices (which it believes to be optimal of course). The humans’ second and third choices further decreases their chances of victory (according to the agent), indicated by the green win probability. However, for the human team’s last two choices, OpenAI Five agrees they were optimal, as can be seen by the win probability remaining constant (even though choice 4, Riki, is a character very differently played by humans and by OpenAI Five).

win probability from the first few frames of every lineup. Given these precomputed win probabilities, we apply a dynamic programming algorithm to draft the best hero available on each turn. Results of this approach in a web-based drafting program that we have built can be seen on Figure 12.

In addition to building a hero selection tool, we also learned about our agent’s preferences from this exercise. In many ways OpenAI Five’s preferences match human player’s preferences such as placing a high value (within this pool) on the hero Sniper. In other ways it does not agree with typical human knowledge, for example it places low value on Earthshaker. Our agent had trouble dealing with geometry of this hero’s “Fissure” skill, making this hero worse than others in training rollouts.

Another interesting tidbit is that at the very start of the draft, before any heroes are picked, OpenAI Five believes that the Radiant team has a 54% win chance (if picking first in the draft) or 53% (if picking second). Our agent’s higher estimate for the Radiant side over the Dire agrees with conventional wisdom within the Dota 2 community. Of course, this likely depends on the set of heroes available.

E Observation Space

At each time step one of our heroes observes $\sim 16,000$ inputs about the game state (mostly real numbers with some integer categorical data as well). See Figure 14 for a schematic outline of our observation space and Table 4 for a full listing of the observations.

Instead of using the pixels on the screen, we approximate the information available to a human player in a set of data arrays. This approximation is imperfect; there are small pieces of information which humans can gain access to which we have not encoded in the observations. On the flip side,

Global data	22	Per-hero add'l (10 heroes)	25	Per-modifier (10 heroes x 10 modifiers & 179 non-heroes x 2 modifiers)	2
time since game started	1	is currently alive?	1	remaining duration	1
is it day or night?		number of deaths	1	stack count	1
time to next day/night change	2	hero currently in sight?		modifier name	1
time to next spawn: creep, neutral, bounty, runes	4	time since this hero last seen	2	Per-item (10 heroes x 16 items)	13
time since seen enemy courier is that > 40 seconds? ^a	2	hero currently teleporting?		location one-hot (inventory/backpack/stash)	3
min&max time to Rosh spawn	2	if so, target coordinates (x, y)	4	charges	1
Roshan's current max hp	1	time they've been channeling	4	is on cooldown?	
is Roshan definitely alive?	1	respawn time	1	cooldown time	2
is Roshan definitely dead?	1	current gold (allies only)	1	is disabled by recent swap?	
Next Roshan drops cheese?	1	level	1	item swap cooldown	2
Next Roshan drops refresher?	1	mana: max, current, & regen	3	toggled state	1
Roshan health randomization ^b	1	health regen rate	1	special Power Treads one-hot (str/agi/int/none)	4
Glyph cooldown (both teams)	2	magic resistance	1	item name	1
Stock counts ^c	4	strength, agility, intelligence	3	Per-ability (10 heroes x 6 abilities)	7
Per-unit (189 units)	43	currently invisible?	1	cooldown time	1
position (x, y, z)	3	is using ability?	1	in use?	1
facing angle (cos, sin)	2	is using ability?	1	castable	1
currently attacking? ^e		# allied/enemy creeps/heroes in line btwn me and this hero ^e	4	Level 1/2/3/4 unlocked? ^d	4
time since last attack ^d	2	Per-allied-hero additional (5 allied heroes)	211	ability name	1
max health		Scripted purchasing settings ^b	7	Per-pickup (6 pickups)	15
last 16 timesteps' hit points	17	Buyback: has?, cost, cooldown	3	status one-hot (present/not present/unknown)	3
attack damage, attack speed	2	Empty inventory & backpack slots	2	location (x, y)	2
physical resistance	1	Lane Assignments ^b	3	distance from all 10 heroes	10
invulnerable due to glyph?		Flattened nearby terrain: 14x14 grid of passable/impassable?	196	pickup name	1
glyph timer	2	scripted build id		Minimap (10 tiles x 10 tiles)	9
movement speed	1	next item to purchase ^b	2	fraction of tile visible	1
on my team? neutral?	2	Nearby map (8x8)^e	6	# allied & enemy creeps	2
animation cycle time	1	terrain: elevation, passable?	2	# allied & enemy wards	2
eta of incoming ranged & tower creep projectile (if any)		allied & enemy creep density	2	# enemy heroes	1
# melee creeps atking this unit ^d	3	area of effect spells in effect. ^f	2	cell (x, y, id)	3
[Shrine only] shrine cooldown	1	area of effect spells in effect. ^f	2		
vector to me (dx, dy, length) ^e	3	Previous Sampled Action^e	310		
am I attacking this unit? ^e		Offset? (Regular, Caster, Ward)	3x2x9		
is this unit attacking me? ^{d,e}		Unit Target's Embedding	128		
eta projectile from unit to me ^e	3	Primary Action's Embedding	128		
unit type	1				
current animation	1				

^a These observations are leftover from an early version of Five which played a restricted 1v1 version of the game. They are likely obsolete and not needed, but this was not tested.

^b These observations are about our per-game randomizations. See Appendix O.

^c For items: gem, smoke of deceit, observer ward, infused raindrop.

^d Observations are not visible per-se, but can be estimated. We use scripted logic to estimate them from visible observations.

^e These observations (only) are different for the five different heroes on the team.

^f This observation appears twice, and serves as an example of the difficulties of surgery. Although this is a categorical input, we began by treating it as a float input to save on engineering work (this observation is unlikely to be very important). Later the time came to upgrade it to a properly embedded categorical input, but our surgery tools do not support removing existing observations. Hence we added the new observation, but were forced to leave the deprecated observation as well.

Table 4: Full Observation Space: All observations OpenAI Five receives at each time step. Blue rows are categorical data. Entries with a question mark are boolean observations (only take values 0 or 1 but treated as floats otherwise). The bulk of the observations are per-unit observations, observed for each of 189 units: heroes (5), creeps (30), buildings (21), wards (30), and courier (1) for each team, plus 15 neutrals. If the number of visible units in a category is less than the allotted number, the rest are padded with zeroes. If more, we observe only the units closest to allied heroes. Units in fog of war are not observed. When enemy heroes are in fog of war, we reuse the observation from the last time step when the unit was visible.

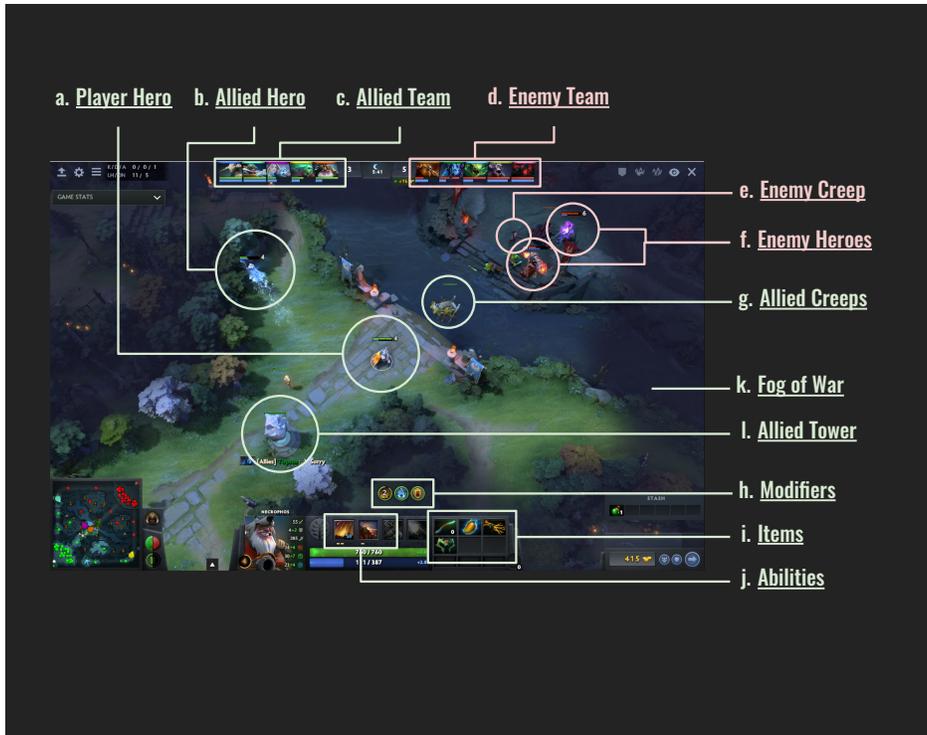


Figure 13: Dota 2’s human “Observation Space”

while we were careful to ensure that all the information available to the model is also available to a human, the model does get to see *all* the information available simultaneously every time step, whereas a human needs to click into various menus and options to get that data. Although these discrepancies are a limitation, we do not believe they meaningfully detract from our ability to benchmark against human players.

Humans observe the game via a rendered screen, depicted in Figure 13. OpenAI Five uses a more semantic observation space than this for two reasons: First, because our goal is to study strategic planning and gameplay rather than focus on visual processing. Second, it is infeasible for us to render each frame to pixels in all training games; this would multiply the computation resources required for the project manyfold.

All float observations (including booleans which are treated as floats that happen to take values 0 or 1) are normalized before feeding into the neural network. For each observation, we keep a running mean and standard deviation of all data ever observed; at each timestep we subtract the mean and divide by the st dev, clipping the final result to be within $(-5, 5)$.

F Action Space

Dota 2 is usually controlled using a mouse and keyboard. The majority of the actions involve a high-level command (attack, use a certain spell, or activate a certain item), along with a target (which might be an enemy unit for an attack, or a spot on the map for a movement). For that reason we represent the action our agent can choose at each timestep as a single *primary action*

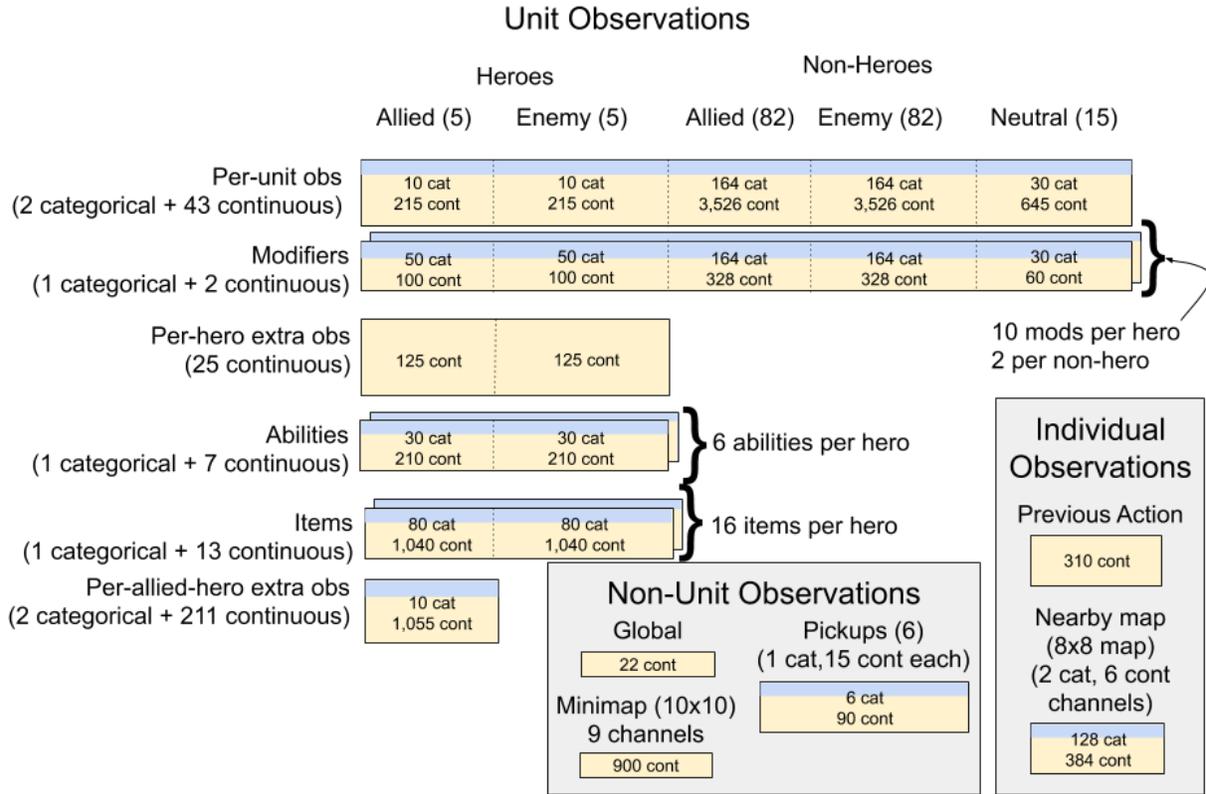


Figure 14: **Observation Space Overview:** The arrays that OpenAI Five observes at each timestep. Most of OpenAI Five’s observations are unit-centered; for 189 different units on the map, we observe a set of basic properties. These units are grouped along the top of the figure. We observe some data about all units, some extra data about the primary units (the heroes), and even more data about the heroes on our team. A few observations are not tied to any unit. Finally, two observations having to do with hero control (terrain near me, and my previous action) are only observed about the individual hero that this LSTM replica operates. In this diagram blue bands represent categorical data and yellow bands represent continuous or boolean data; most entities (units, modifiers, abilities, items, and pickups), have some of each. Each piece of the figure summarizes the total dimensionality of that portion of the input. All together, an OpenAI Five hero observes 1,200 categorical values and 14,534 continuous/boolean values.



(a) **Delay:** An integer from 0 to 3 indicating which frame during the next frameskip to take the action on (see Appendix L). If 0, the action will be taken immediately when the game engine processes this time step; if 3, the action will be taken on the last game frame before the next policy observation. This parameter is never ignored.

(b) **Unit Selection:** One of the 189 visible units in the observation. For actions and abilities which target units, either enemy units or friendly units. For many actions, some of the possible unit targets will be invalid; attempting an action with an invalid target results in a noop.

(c) **Offset:** A 2D (X, Y) coordinate indicating a spatial offset, used for abilities which target a location on the map. The offset is interpreted relative to the caster or the unit selected by the Unit Selection parameter, depending on the ability. Both X and Y are discrete integer outputs ranging from -4 to +4 inclusive, producing a grid of 81 possible coordinate pairs.

Figure 15: **Action Parameters**

along with a number of *parameter actions*.

The number of primary actions available varies from time step to time step, averaging 8.1 in the games against OG. The primary actions available at a given time include universal actions like noop, move, attack, and others; use or activate one of the hero’s spells; use or activate one of the hero’s items; situational actions such as Buyback (if dead), Shrine (if near a shrine), or Purchase (if near a shop); and more. For many of the actions we wrote simple action filters, which determine whether the action is available; these check if there is a valid target nearby, if the ability/item is on cooldown, etc. At each timestep we restrict the set of available actions using these filters and present the final choices to the model.

In addition to a primary action, the model chooses action parameters. At each timestep the model outputs a value for each of them; depending on the primary action, some of them are read and others ignored (when optimizing, we mask out the ignored ones since their gradients would be pure noise). There are 3 parameter outputs, Delay (4 dim), unit selection (189 dim), and offset (81 dim), described in Figure 15.

All together this produces a combined factorized action space size of up to $30 \times 4 \times 189 \times 81 = 1,837,080$ dimensions (30 being the maximum number of primary actions we support). This number ignores the fact that the number of primary actions is usually much lower; some parameters are masked depending on the primary action; and some parameter combinations are invalid and those actions are treated as no-ops.

To get a better picture, we looked at actual data from the two games played against Team OG, and simply counted number of available actions at each step. The average number of available

Action Target Type	Example	Parameters
No Target	Power Treads	Delay
Point Target	Move	Delay, Offset (Caster)
Unit Target	Attack	Delay, Unit Selection (Regular)
Unit Offset Target	Sniper’s Shrapnel	Delay, Unit Selection (Regular), Offset (Regular)
Teleport Target	Town Portal Scroll	Delay, Unit Selection (Teleport), Offset (Regular)
Ward Target	Place Observer Ward	Delay, Offset (Ward)

Table 5: **Action Target Types**

actions varies significantly across heroes, as different heroes have different numbers spells and items with larger parameter counts. Across the two games the average number of actions for a hero varied from 8,000 to 80,000.

Unit Selection and Offset are actually implemented within the model as several different, mutually exclusive parameters depending on the primary action. For Unit Selection, we found that using a single output head caused that head to learn very well to target tactical spells and abilities. One ability called “teleport,” however, is significantly different from all the others — rather than being used in a tactical fight, it is used to strategically reposition units across the map. Because the action is much more rare, the learning signal for targeting this ability would be drowned out if we used a single model output head for both. For this reason the model outputs a normal Unit Selection parameter and a separate Teleport Selection parameter, and one or the other is used depending on the primary action. Similarly, the Offset parameter is split into “Regular Offset,” “Caster Offset” (for actions which only make sense offset from the caster), and “Ward Placement Offset” (for the rare action of placing observer wards).

We categorize all primary actions into 6 “Action target types” which determines which parameters the action uses, listed in Table 5.

F.1 Scripted Actions

Not all actions that a human takes in a game of Dota 2 are controlled by our RL agent. Some of the actions are *scripted*, meaning that we have written a rudimentary rules-based system to handle these decisions. Most of these are for historical reasons — at the start of the project we gave the model control over a small set of the actions, and we gradually expanded it over time. Each additional action that we remove from the scripted logic and hand to the model’s control gives the RL system a higher potential skill cap, but comes with a cost measured in engineering effort to set it up and risks associated with learning and exploration. Indeed even when adding these new actions gradually and systematically, we occasionally encountered instabilities; for example the agent might quickly learn never to take a new action (and thus fail to explore the small fraction of circumstances where that action helps), and thus moreover fail to learn (or unlearn) the dependent parts of the gameplay which require competent use of the new action.

In the end there were still several systems that we had not yet removed from the scripted logic by the time the agent reached superhuman performance. While we believe the agent could ultimately perform better if these actions were not scripted, we saw no reason to do remove the scripting because superhuman performance had already been achieved. The full set of remaining scripted actions is:

1. **Ability Builds:** Each hero has four spell abilities. Over the course of the game, a player can choose which of these to “level up,” making that particular skill more powerful. For these, in evaluation games we follow a fixed schedule (improve ability X at level 1, then Y at level 2, then Z at level 3, etc). In training, we randomize around this fixed script somewhat to ensure the model is robust to the opponent choosing a different schedule.
2. **Item Purchasing:** As a hero gains gold, they can purchase items. We divide items into *consumables* — items which are consumed for a one-time benefit such as healing — and everything else. For consumables, we use a simple logic which ensures that the agent always has a certain set of consumables; when the agent uses one up, we then purchase a new one. After a certain time in the game, we stop purchasing consumables. For the non-consumables we use a system similar to the ability builds - we follow a fixed schedule (first build X, then Y, then Z, etc). Again at training time we randomly perturb these builds to ensure robustness to opponents using different items.¹¹
3. **Item Swap:** Each player can choose 6 of the items they hold to keep in their “inventory” where they are actively usable, leaving up to 3 inactive items in their “backpack.” Instead of letting the model control this, we use a heuristic which approximately keeps the most valuable items in the inventory.
4. **Courier Control:** Each side has a single “Courier” unit which cannot fight but can carry items from the shop to the player which purchased them. We use a state-machine based logic to control this character.

G Reward Weights

Our agent’s ultimate goal is to win the game. In order to simplify the credit assignment problem (the task of figuring out which of the many actions the agent took during the game led to the final positive or negative reward), we use a more detailed reward function. Our shaped reward is modeled loosely after potential-based shaping functions [58], though the guarantees therein do not apply here. We give the agent reward (or penalty) for a set of actions which humans playing the game generally agree to be good (gaining resources, killing enemies, etc).

All the results that we reward can be found in Table 6, with the amount of the reward. Some are given to every hero on the team (“Team”) and some just to the hero who took the action “Solo.” Note that this means that when team spirit is 1.0, the total amount of reward is five times higher for “Team” rewards than “Solo” rewards.

In addition to the set of actions rewarded and their weights, our reward function contains 3 other pieces:

- **Zero sum:** The game is zero sum (only one team can win), everything that benefits one team necessarily hurts the other team. We ensure that all our rewards are zero-sum, by subtracting from each hero’s reward the average of the enemies’ rewards.

¹¹This randomization is done randomly deleting items from the build order and randomly inserting new items sampled from the distribution of which items that hero usually buys in human games. This is the only place in our system which relies on data from human games.

Name	Reward	Heroes	Description
Win	5	Team	
Hero Death	-1	Solo	
Courier Death	-2	Team	
XP Gained	0.002	Solo	
Gold Gained	0.006	Solo	For each unit of gold gained. Reward is not lost when the gold is spent or lost.
Gold Spent	0.0006	Solo	Per unit of gold spent on items without using courier.
Health Changed	2	Solo	Measured as a fraction of hero's max health. [‡]
Mana Changed	0.75	Solo	Measured as a fraction of hero's max mana.
Killed Hero	-0.6	Solo	For killing an enemy hero. The gold and experience reward is very high, so this reduces the total reward for killing enemies.
Last Hit	-0.16	Solo	The gold and experience reward is very high, so this reduces the total reward for last hit to ~ 0.4 .
Deny	0.15	Solo	
Gained Aegis	5	Team	
Ancient HP Change	5	Team	Measured as a fraction of ancient's max health.
Megas Unlocked	4	Team	
T1 Tower [*]	2.25	Team	
T2 Tower [*]	3	Team	
T3 Tower [*]	4.5	Team	
T4 Tower [*]	2.25	Team	
Shrine [*]	2.25	Team	
Barracks [*]	6	Team	
Lane Assign [†]	-0.15	Solo	Per second in wrong lane.

^{*} For buildings, two-thirds of the reward is earned linearly as the building loses health, and one-third is earned as a lump sum when it dies.

[†] See item O.2.

[‡] Hero's health is quartically interpolated between 0 (dead) and 1 (full health); health at fraction x of full health is worth $(x + 1 - (1 - x)^4) / 2$. This function was not tuned; it was set once and then untouched for the duration of the project.

Table 6: Shaped Reward Weights

- **Game time weighting:** Each player’s “power” increases dramatically over the course of a game of Dota 2. A character who struggled to kill a single weak creep early in the game can often kill many at once with a single stroke by the end of the game. This means that the end of the game simply produces more rewards in total (positive or negative). If we do not account for this, the learning procedure focuses entirely on the later stages of the game and ignores the earlier stages because they have less total reward magnitude. We use a simple renormalization to deal with this, multiplying all rewards other than the win/loss reward by a factor which decays exponentially over the course of the game. Each reward ρ_i earned a time T since the game began is scaled:

$$\rho_i \leftarrow \rho_i \times 0.6^{(T/10 \text{ mins})} \tag{11}$$

- **Team Spirit:** Because we have multiple agents on one team, we have an additional dimension to the credit assignment problem, where the agents need learn which of the five agent’s behavior cause some positive outcome. The partial rewards defined in Table 6 are an attempt to make the credit assignment easier, but they may backfire and in fact add more variance if an agent receives reward when a *different* agent takes a good action. To attempt dealing with this, we have introduced *team spirit*. It measures how much agents on the team share in the spoils of their teammates. If each hero earns raw individual reward ρ_i , then we compute the hero’s final reward r_i as follows:

$$r_i = (1 - \tau)\rho_i + \tau\bar{\rho} \tag{12}$$

with scalar $\bar{\rho}$ being equal to mean of ρ . If team spirit is 0, then it’s every hero for themselves; each hero only receives reward for their own actions $r_i = \rho_i$. If team spirit is 1, then every reward is split equally among all five heroes; $r_i = \bar{\rho}$. For a team spirit τ in between, team spirit-adjusted rewards are linearly interpolated between the two.

Ultimately we care about optimizing for team spirit $\tau = 1$; we want the actions to be chosen to optimize the success of the entire team. However we find that lower team spirit reduces gradient variance in early training, ensuring that agents receive clearer reward for advancing their mechanical and tactical ability to participate in fights individually. See Appendix O for an ablation of this method.

We ran a small-scale ablation with partial reward weights disabled (see Figure 16). Surprisingly, the model learned to play well enough to beat a hand-coded scripted agent consistently, though with a large penalty to sample efficiency relative to the shaped reward baseline. From watching these games, it appears that this policy does not play as effectively at the beginning of the game, but has learned to coordinate fights nearer to the end of the game. Investigating the tradeoffs and benefits of sparse rewards is an interesting direction for future work.

H Neural Network Architecture

A simplified diagram of the joint policy and value network is shown in the main text in Figure 1. The combined policy + value network uses 158,502,815 parameters (in the final version).

The policy network is designed to receive observations from our bot-API observation space, and interact with the game using a rich factorized action space. These structured observation and action spaces heavily inform the neural network architecture used. We use five replica neural networks,

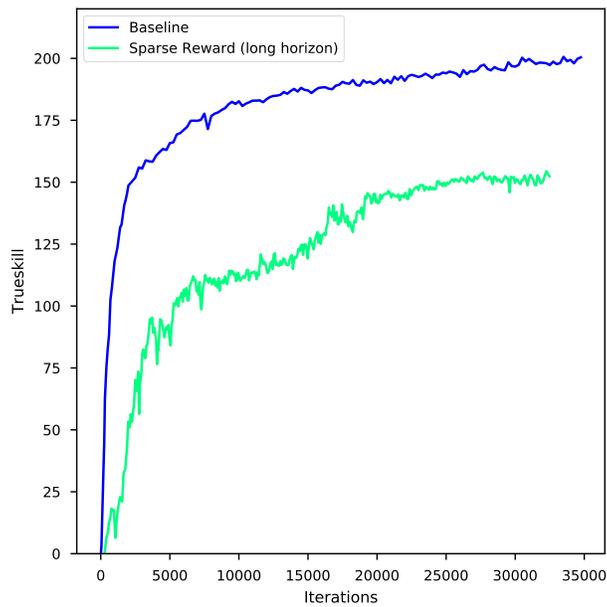
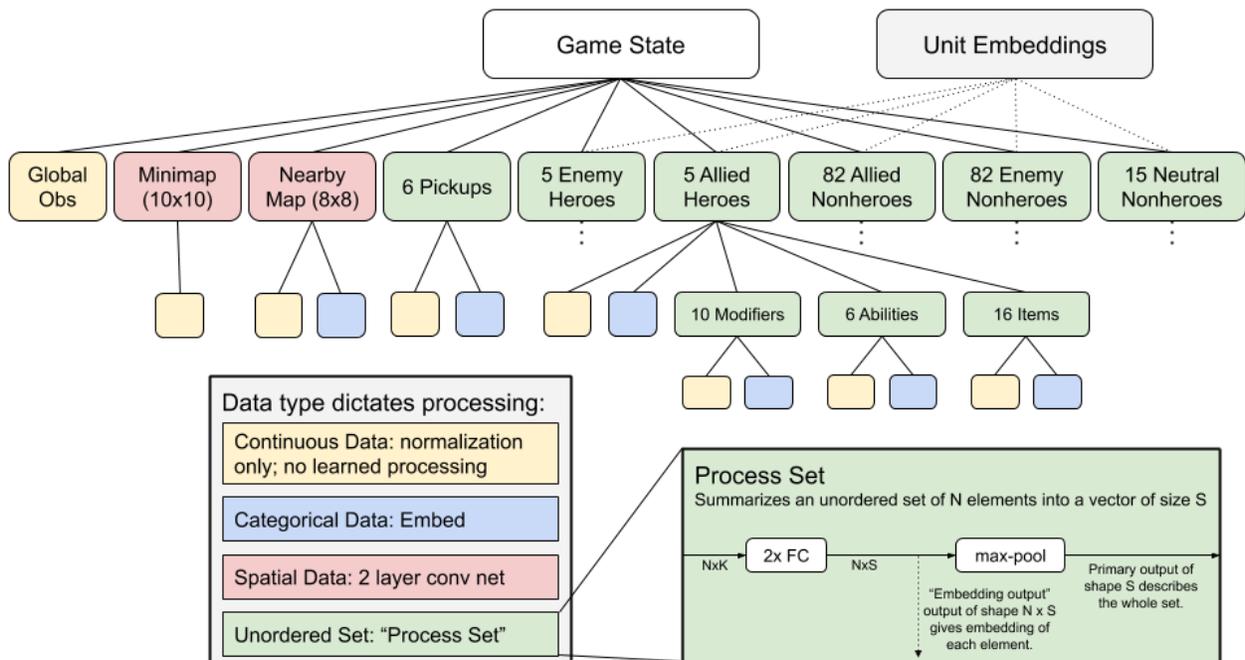
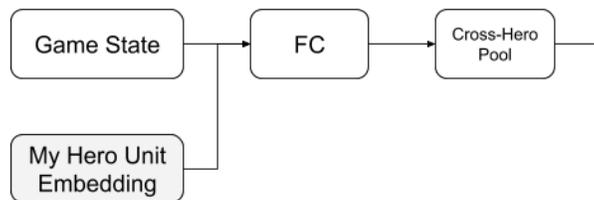


Figure 16: **Sparse rewards in Dota 2:** TrueSkill over the course of training for experiments run with 0-1 loss only. For the sparse reward run horizon was set to 1 hour ($\gamma = 0.99996$) (versus 180 seconds for the baseline run). The baseline otherwise uses identical settings and hyperparameters including our shaped reward. The sparse reward succeeds at reaching TrueSkill 155; for reference, a hand-coded scripted agent reaches TrueSkill 100.



(a) **Flattening the observation space:** First we process the complicated observation space into a single vector. The observation space has a tree structure; the full game state has various attributes such as global continuous data and a set of allied heroes. Each allied hero in turn has a set of abilities, a set of modifiers, etc. We process each node in the tree according to its data type. For example for spatial data, we concatenate the data within each cell and then apply a 2 layer conv net. For unordered sets, a common feature of our observations, we use a “Process Set” module. Weights in the Process Set module for processing abilities/items/modifiers are shared across allied and enemy heroes; weights for processing modifiers are shared across allied/enemy/neutral nonheroes. In addition to the main Game State observation, we extract the the Unit Embeddings from the “embedding output” of the units’ process sets, for use in the output (see Figure 18).



(b) **Preparing for LSTM:** In order to tell each LSTM which of the team’s heroes it controls, we append the controlled hero’s Unit Embedding from the Unit Embeddings output of Figure 17a to the Game State vector. Almost all of the inputs are the same for each of the five replica LSTMs (the only differences are the nearby map, previous action, and a very small fraction of the observations for each unit). In order to allow each replica to respond to the non-identical inputs of other replicas if needed, we add a “cross-hero pool” operation, in which we maxpool the first 25% of the vector across the five replica networks.

Figure 17: **Observation processing in OpenAI Five**

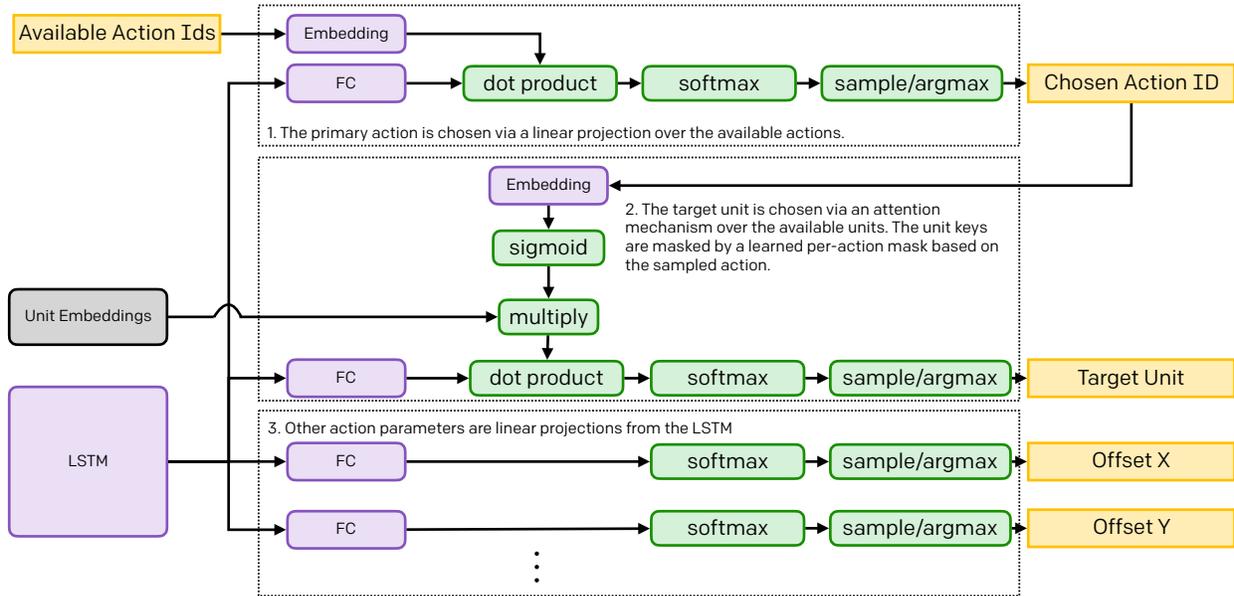


Figure 18: The hidden state of the LSTM and unit embeddings are used to parameterize the actions.

each responsible for the observations and actions of one of the heroes in the team. At a high level, this network consists of three parts: first the observations are processed and pooled into a single vector summarizing the state (see Figure 17), then that is processed by a single-layer large LSTM, then the outputs of that LSTM are projected to produce outputs using linear projections (see Figure 18).

To provide the full details, we should clarify that Figure 1 is a slight over-simplification in three ways:

1. In practice the Observation Processing portion of the model is also cloned 5 times for the five different heroes. The weights are identical and the observations are nearly identical — but there are a handful of derived features which are different for each replica (such as “distance to me” for each unit; see Table 4 for the list of observations that vary). Thus the five replicas produce nearly identical, but perhaps not entirely identical, LSTM inputs. These non-identical features form a small portion of the observation space, and were not ablated; it is possible that they are not needed at all.
2. The “Flattened Observation” and “Hero Embedding” are processed before being sent into the LSTM (see Figure 17b) by a fully-connected layer and a “cross-hero pool” operation, to ensure that the non-identical observations can be used by other members of the team if needed.
3. The “Unit Embeddings” from the observation processing are carried along beside the LSTM, and used by the action heads to choose a unit to target (see Figure 18).

In addition to the action logits, the value function is computed as another linear projection of the LSTM state. Thus our value function and action policy share a network and share gradients.

I Human Games

See Table 7 for a listing of the games OpenAI Five played against high-profile teams.

J TrueSkill: Evaluating a Dota 2 Agent Automatically

We use the TrueSkill [27] rating system to evaluate our agents.

We first establish a pool of many reference agents of known skill. We evaluate the reference agents’ TrueSkill by playing many games between the the various reference agents, and using the outcome of the games to compute a TrueSkill for each agent. Our TrueSkill environment use the parameters $\sigma = 25/3$, $\beta = \sigma/2$, $\tau = 0.0$, `draw_probability=0.02`. Reference agents’ μ are aligned so that an agent playing randomly has $\mu = 0$. A hand-crafted scripted agent which we wrote, which can defeat beginners but not amateur players, has TrueSkill around 105.

During our experiments we continually added new reference agents as our agent “outgrew” the existing ones. For all results in this work, however, use a single reference agent pool containing mostly agents from OpenAI Five’s training history along with some other smaller experiments at the lower end. The 83 reference agents range in TrueSkill from 0 (random play) to 254 (the version that beat the world champions).

To evaluate a training run during training, a dedicated set of computers continually download the latest agent parameters and plays games between the latest trained agent and the reference agents. We attempt to only play games against reference agents that are nearby in skill in order to gain maximally useful information; we avoid playing agents more than 10 TrueSkill points away (corresponding to a winrate less than 15% or more than 85%). When a game finishes, we use the TrueSkill algorithm to update the test agent’s TrueSkill, but treat the reference agent’s TrueSkill as a constant. After 750 games have been reported, we log that version’s TrueSkill and move on to the new current version. New agents are initialized with μ equal to the final μ of the previous agent. This system gives us updates approximately once every two hours during running experiments.

One difficulty in using TrueSkill across a long training experiment was maintaining consistent metrics with a changing environment. Two agents that were trained on different game versions must ultimately play on a single version of the game, which will result in an inherent advantage for the agent that trained on it. Older agents had their code upgraded in order to always be compatible with the newest version, but this still leads to metric inflation for newer agents who got to train on the same code they are evaluated on. This included any updates to the hero pool (adding new heroes that old agents didn’t train with), game client updates or balancing changes, and adding any new actions (using a particular consumable or item differently).

K Dota 2 Gym Environment

K.1 Data flow between the training environment and Dota 2

Dota 2 includes a scripting API designed for building bots. The provided API is exposed through Lua and has methods for querying the visible state of the game as well as submitting actions for bots to take. Parts of the map that are out of sight are considered to be in the fog of war and cannot be queried through the scripting API, which prevents us from accidentally “cheating” by observing anything a human player would not be able to see (although see Appendix Q).

Opponent	Result	Duration	Version	Restrictions
June 6, 2018 - Internal Event				
Internal team	win	15:15 (surr)	7.13	Mirror match, multiple couriers, no invis
Internal team	win	20:51	7.13	Mirror match, multiple couriers, no invis
Audience team	win	31:33	7.13	Mirror match, multiple couriers, no invis
Audience team	win	23:33 (surr)	7.13	Mirror match, multiple couriers, no invis
August 5, 2018 - Benchmark				
Caster team	win	21:38 (surr)	7.16	Drafted, multiple couriers
Caster team	win	24:56 (surr)	7.16	Drafted, multiple couriers
Caster team	lose	35:47	7.16	Audience draft, multiple couriers
August 9, 2018 - Private eval				
Team Secret	win	17:00 (surr)	7.16	Drafted, multiple couriers
Team Secret	lose	48:46	7.16	Drafted, multiple couriers
Team Secret	lose	38:55	7.16	Drafted, multiple couriers
August 22-23, 2018 - The International				
Pain Gaming	lose	52:29	7.19	Pre-set lineup
Chinese Legends	lose	45:44	7.19	Pre-set lineup
October 5, 2018 - Private eval				
Team Lithium	win	48:57	7.19	TI pre-set lineup
Team Lithium	win	48:16	7.19	TI pre-set lineup
Team Lithium	win	31:33	7.19	Drafted
January 16, 2019 - Private eval				
SG Esports	win	24:29 (surr)	7.19	TI pre-set lineup
SG Esports	win	25:08 (surr)	7.19	Drafted
SG Esports	win	27:36 (surr)	7.20	Mirror match
SG Esports	win	25:30 (surr)	7.20	Mirror match
February 1, 2019 - Private eval				
Alliance	win	17:11	7.20d	Drafted
Alliance	win	31:33	7.20d	Drafted
Alliance	win	28:16	7.20d	Reverse drafted
April 13, 2019 - OpenAI Five Finals				
OG	win	38:18	7.21d	Drafted
OG	win	20:51	7.21d	Drafted

Table 7: Major matches of OpenAI Five against high-skill human players.

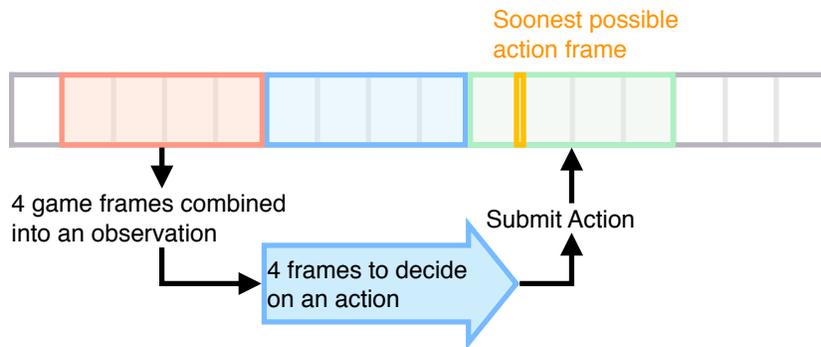


Figure 19: **Reaction Time:** OpenAI Five observes four frames bundled together, so any surprising new information will become available at a random frame in the red region. The model then processes the observation in parallel while the game engine runs forward four more frames. The soonest it can submit an action based on the red observations is marked in yellow. This is between 5 and 8 frames (167-267ms) after the surprising event.

We designed our Dota 2 environment to behave like a standard OpenAI Gym environment[59]. This standard respects an API contract where a `step` method takes action parameters and returns an observation from the next state of the environment. To send actions to Dota 2, we implemented a helper process in Go that we load into Dota 2 through an attached debugger that exposes a gRPC server. This gRPC server implements methods to configure a game and perform an environment step. By running the game with an embedded server, we are able to communicate with it over the network from any remote process.

When the `step` method is called in the gRPC server, it gets dispatched to the Lua code and then the method blocks until an observation arrives back from Lua to be returned to the caller. In parallel, the Dota 2 engine runs our Lua code on every step, sending the current game state observation¹² to the gRPC server and waiting for it to return the current action. The game blocks until an action is available. These two parallel processes end up meeting in the middle, exchanging actions from gRPC in return for observations from Lua. Go was chosen to make this architecture easy to implement through its channels feature.

Putting the game environment behind a gRPC server allowed us to package the game into a Docker image and easily run many isolated game instances per machine. It also allowed us to easily setup, reset, and use the environment from anywhere where Docker is running. This design choice significantly improved researcher productivity when iterating on and debugging this system.

L Reaction time

The Dota 2 game engine runs at 30 steps per second so in theory a bot could submit an action every 33ms. Both to speed up our game execution and in order to bring reactions of our model closer

¹²Originally the Lua scripting API was used to iterate and gather the visible game state, however this was somewhat slow and our final system used an all-in-one game state collection method that was added through cooperation with Valve

to the human scale we downsample to every 4th frame, which we call *frameskip*. This yields an effective observation and action rate of 7.5 frames per second. To allow the model to take precisely timed actions, the action space includes a “delay” which indicates which frame during the frameskip the model wants this action to evaluate on. Thus the model can still take actions at a particular frame if so desired, although in practice we found that the model did not learn to do this and simply taking the action at the start of the frameskip was better.

Moreover, we reduce our computational requirements by allowing the game and the machine learning model to run concurrently by asynchronously issuing actions with an *action offset*. When the model receives an observation at time T , rather than making the game engine wait for the model to produce an action at time T , we let the game engine carry on running until it produces an observation at time $T + 1$. The game engine then sends the observation at time $T + 1$ to the model, and by this time the model has produced its action choice based on the observation at time T . In this way the action which the model takes at time $T + 1$ is based upon the observation at time T . In exchange for this penalty in available “reaction time,” we are able to utilize our compute resources much more efficiently by preventing the two major computations from blocking one another (see Figure 19).

Taken together, these effects mean that the agent can react to new information with a reaction time randomly distributed between 5 and 8 frames (167ms to 267ms), depending on when during the frameskip the new information happens to occur. For comparison, human reaction time has been measured at 250ms in controlled experimental settings[26]. This is likely an underestimate of reaction time during a Dota game.

M Scale and Data Quality Ablation Details

As shown in Figure 5 of the main text, we studied several key ingredients of RL at this scale, and learned important lessons which we conjecture should generalize beyond this environment. In this section we explain the details of these experiments.

Training runs the size of OpenAI Five are expensive; running a scan of 4 different variants would be prohibitively expensive. For this reason we use the normal Dota 2 environment, simply using a batch size 8x smaller than Rerun (which itself was 2-3 times smaller than OpenAI Five). See Figure 20 for an estimate of the variation in these training runs.

Throughout the following sections we scan over various parameters of the experimental setup and monitor the results in terms of TrueSkill (see Appendix J) and speedup (see Equation 2).

Our the uncertainty on speedup comes from uncertainty in both the numerator and the denominator. Although we have some understanding in the variance in the number of iterations for a baseline to reach each TrueSkill (see Figure 20), we do not have the luxury of multiple runs of every experiment. Instead, we use as proxy for the uncertainty on the number of iterations to reach TrueSkill T , the number of iterations to reach to reach $T \pm \Delta T$ where ΔT is the variance in TrueSkill across the variations in Figure 20, approximately 2 TrueSkill points. We combine the numerator and denominator uncertainty in quadrature to attain an overall uncertainty for the speedup.

In each experiment the baseline uses hyperparameters given in Appendix C, except as noted.

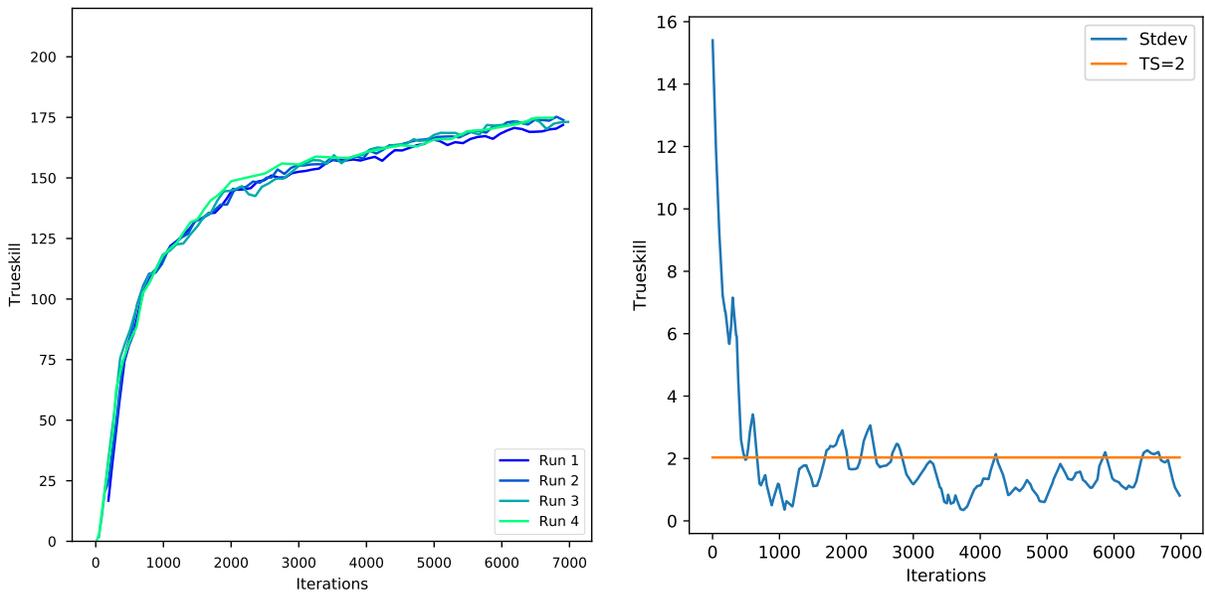


Figure 20: **Variation in 5v5 baseline training:** On the left, the TrueSkill over the course of training for different “baseline” experiments, using identical settings and hyperparameters. On the right, the standard deviation in TrueSkill across four runs. See Appendix C for the hyperparameters used. Although we only have 4 runs, we can estimate that different runs tend to vary by about 2 TrueSkill.

M.1 Batch Size

Training using small mini-batches is a generally accepted trade-off between convergence time and number of optimization steps. However, recent literature on large-scale supervised learning of image classifiers [44–46] explored much larger batch sizes and showed that strong scaling was possible by carefully tuning learning rate and initialization of the neural network. This renewed interest in reducing convergence-time and treating batch-size as a key design parameter also motivated the work of [28], where an analytical tool is derived to estimate a training-time optimal batch size on per task basis by studying the “noise scale” of the gradients.

While existing literature on large-scale training of neural networks had focused on supervised learning, as far as we know using large batch sizes for reinforcement learning was novel when we began the Dota 2 project. These observations were later shown to be consistent with the analytical tools derived in [28]. In this section we demonstrate how large batch-sizes affect optimization time.

Because we average gradients across the pool of optimizer machines, the effective total batch size is given by the product of the number of GPU optimizers with the batch size on each optimizer. We always use the maximum batch size on each optimizer which will fit within the GPU’s memory constraints (120 for our setup). Thus in order to change the overall batch size we increase the number of optimizer GPUs. We increase the size of the other machine pools in the experiment (rollout CPU workers, forward pass GPUs, etc), such that the larger batch size experiment is truly optimizing over more data, not simply reusing the same data more. This means that doubling the batch size causes the experiment to use twice as much computing power in almost all respects. Because we do not have the resources to separately optimize these hyperparameters at each individual batch size,

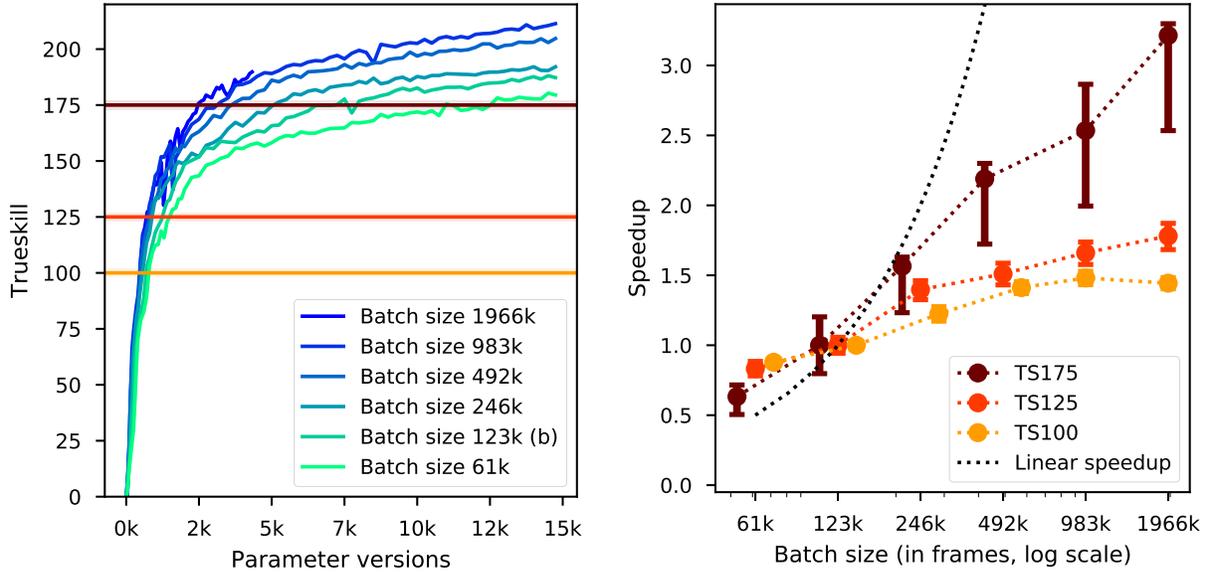


Figure 21: **Effect of batch size on training speed:** (Replicated from main text Figure 5a) TrueSkill over the course of training (see Appendix J) and speedup measured by the rate to attain different TrueSkill thresholds (computed using Equation 2) granted by increasing the batch size. The dotted line indicates perfect linear scaling (using 2x more data gives 2x speedup). Larger batch size significantly speeds up training, but the speedup is sublinear in the resources consumed. Later training (TrueSkill 175) benefits more from increased scale than earlier training (TrueSkill 100). Note that TrueSkill 175 is still quite early in the overall training of OpenAI Five which ultimately reaches above 250 (see Figure 3), so these results are inconclusive about whether large batch size causes linear speedup for the bulk of the training time.

we keep all other hyperparameters fixed to those listed under “baseline” in Table 2.

Results can be seen in Figure 5a, with discussion in the main text.

M.2 Sample Quality — Staleness

In an ideal world, each piece of data in the optimizer would be perfectly on-policy (to obtain unbiased gradients), would be used exactly once and then thrown out (to avoid overfitting), would be from a completely different episode than every other piece of data (to eliminate correlations), and more. Because of our enormous batch size and small learning rate, we hypothesized that loosening the above constraints would not be a large price to pay in exchange for the benefits of asynchronous processing. However, we actually learned that issues like this surrounding data quality can be quite significant. In this and next section we will focus on two of these issues, which we call *staleness* and *sample reuse*.

Early on in the development of our agent we would play the whole game of Dota 2 using single set of parameters, then send this huge package of sample data to optimizers for training. One of the negative effects of this approach was that this would render data stale; the policy parameters which played the start of the game would be an hour old or more, making the gradients estimated from them incorrect. Therefore we have switched to accumulating small amount of training data;

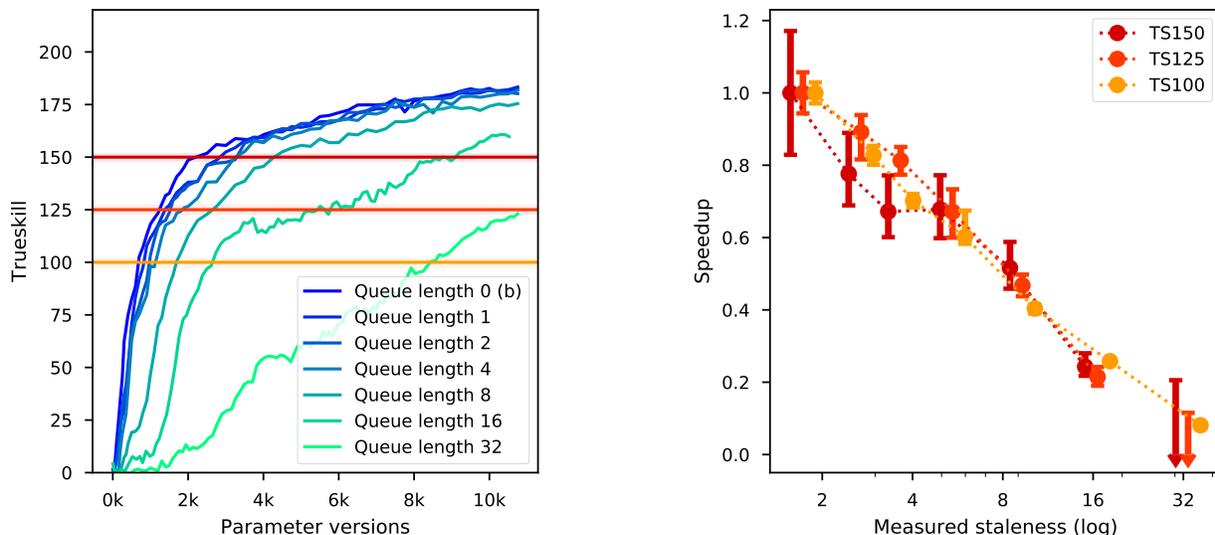


Figure 22: **Effect of Staleness on training speed.** (Replicated from main text Figure 5b) TrueSkill over the course of training (see Appendix J) and speedup measured by the rate to attain different TrueSkill thresholds (computed using Equation 2) granted by increasing Staleness. Increasing staleness of data causes significant losses in training speed.

sending it over to optimizers and updating agent parameters; then continuing with the same game.

In order to generate rollouts with a certain version of the parameters, a long round-trip has to happen (see Figure 2). This new set of parameters is published to the controller, then independently pulled by forward pass machines, which only then will start using this version of parameters to perform forward-passes of our agent. Then some amount of gameplay must be rolled forward and after that the data is finally sent to the optimizers. In the meanwhile, the optimizers have been running on previously-collected data and advanced by some number of new gradient descent steps. In our setup where rollouts send about 30 seconds of gameplay in each chunk, this loop takes 1-2 minutes. Because our learning rate is small, and this is only a few minutes on the scale of a multi-week learning endeavor, one might expect this to be a minor concern — but to the contrary, we observe that this it can be a crucial detail.

In this study we artificially introducing additional delay to see the effect. This is implemented on the rollout workers; instead of sending their data immediately back to the optimizers, they now put it in a queue, and pop data off the end of it to send to the optimizers. Thus the length of the queue determines the amount of artificial staleness introduced. See Figure 23; we observe the desired increase in measured staleness with the length of the queue.

The results can be found in the main text in Figure 5b, and are reproduced in Figure 22. Staleness negatively affects speed of training, and the drop can be quite severe when the staleness is larger than a few versions. For this reason we attempt to keep staleness as low as possible in our experiments.

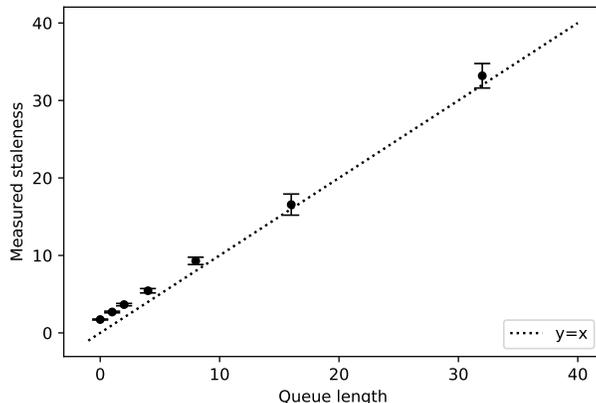


Figure 23: Adding a queue that buffers rollout data on the way to optimizers increases measured staleness in a predictable manner. Error bars indicate the standard deviation of measured staleness as it varied over the course of training due to distributed systems fluctuations.

M.3 Sample Quality — Sampling and Sample Reuse

Our asynchronous training system reuses samples in multiple optimization steps. Each optimizer’s experience buffer is constantly asynchronously collecting data from rollout machines. At each optimization step, a batch of data is sampled from this buffer. The buffer is configured to hold 4096 samples. Our optimizers compute the average sample reuse as the ratio between data arrival and consumption rates:

$$\text{Sample Reuse} \equiv \frac{(\text{samples per batch}) \times (\text{batches per second})}{(\text{experience buffer intake samples per second})} \quad (13)$$

Sample reuse is a function of the round trip time between rollout machines and optimizers, the ratio of rollout machines to optimizers, and other factors, and thus we only approximately hit target values but do not set them exactly. We measure the effect of sample reuse by varying the rate of incoming samples to the optimizers. In practice, the rate of data production from each rollout worker stays relatively stable, so we vary this rate by changing the number of rollout CPU workers and forward pass GPUs while keeping the number of optimizers and everything else fixed.

Our baseline experiment is tuned to have a sample reuse of approximately 1. To measure the effect of sample reuse we reduced the number of rollouts by 2, 4, and 8x to induce higher sample reuse. Additionally we also doubled the number of rollouts for one experiment to investigate the regime where sample reuse is lower than 1. These adjustments yielded sample reuse measurements between 0.57 and 6.3 (see Figure 25). It is important to highlight that adjusting the number of rollouts directly affects the number of simultaneous games being played, which affects the diversity of games that are used for training.

The results can be found in the main text in Figure 5c, and are reproduced in Figure 24. We found that increasing sample reuse causes a significant decrease in performance. As long as the optimizers are reusing data, adding additional rollout workers appears to be a relatively cheap way to accelerate training. CPUs are often easier and cheaper to scale up than GPUs and this can be a significant performance boost in some setups.

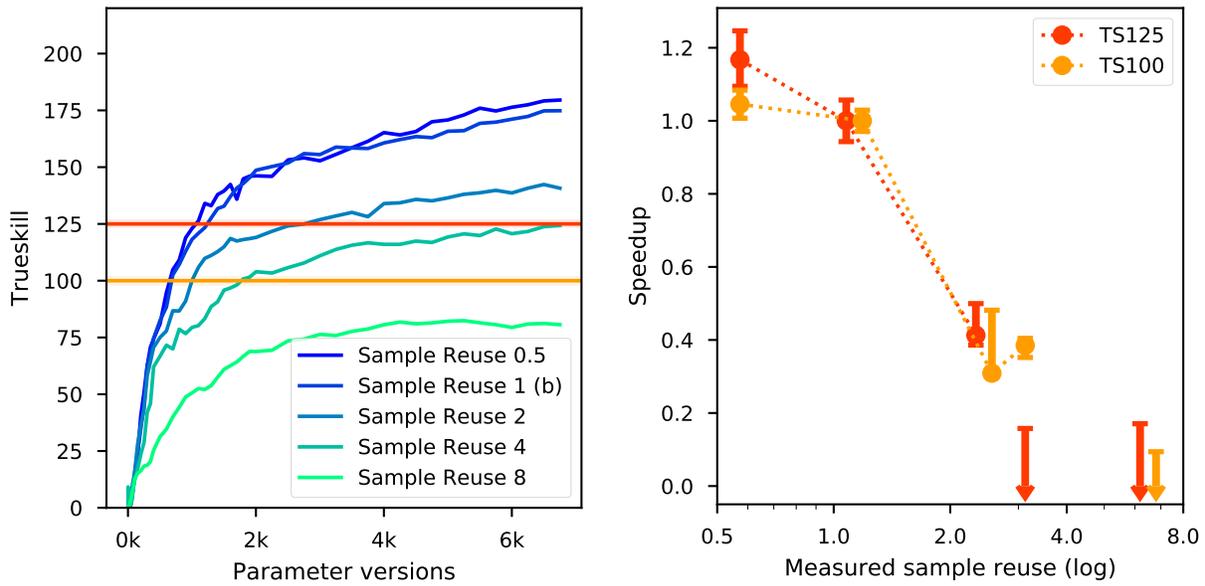


Figure 24: **Effect of Sample Reuse on training speed.** (Replicated from main text Figure 5c) TrueSkill over the course of training (see Appendix J) and speedup measured by the rate to attain different TrueSkill thresholds (computed using Equation 2) granted by increasing Sample Reuse. Increasing sample reuse causes significant slowdowns. In fact, the run with 1/8th as many rollout workers (sample reuse around 6.3), seems to have converged to less than 75 TrueSkill.

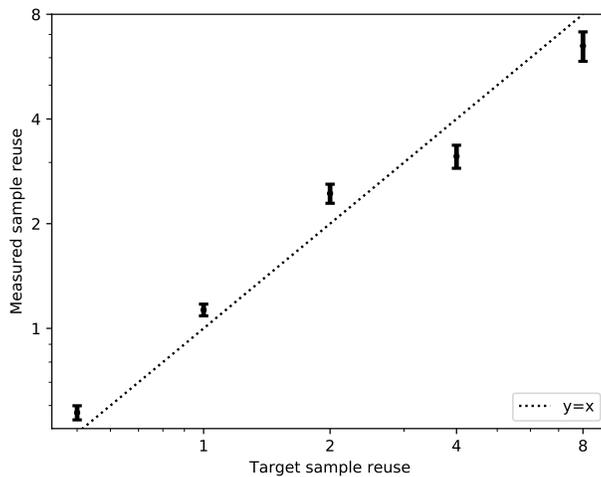


Figure 25: As our target sample reuse increases measured sample reuse increases predictably. Error bars indicate the standard deviation of measured sample reuse as it varied over the course of training.

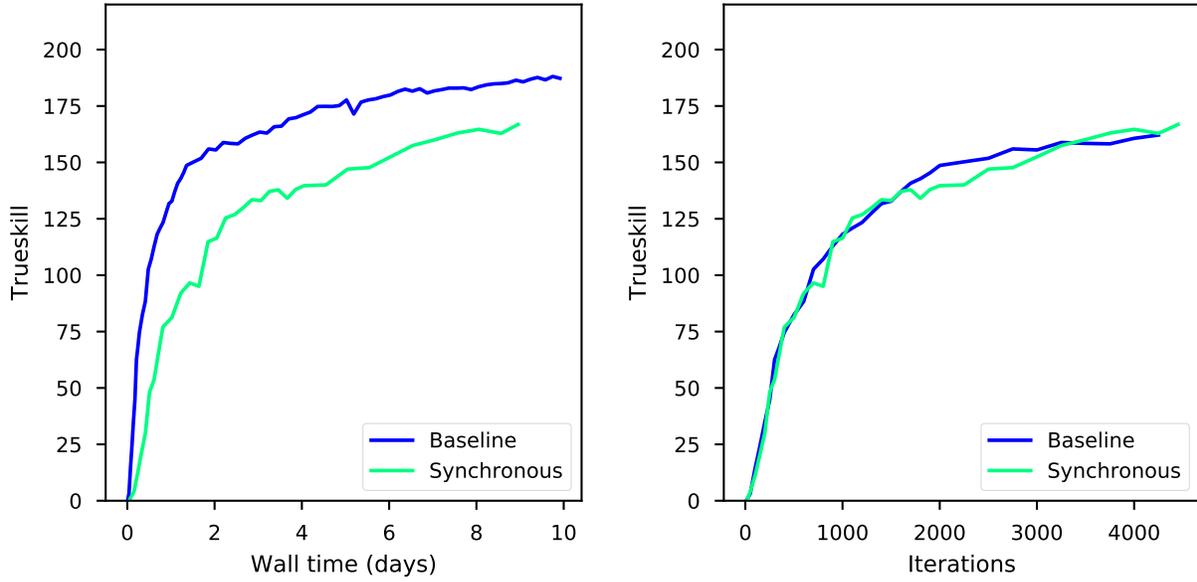


Figure 26: **Asynchronous training:** Plots of TrueSkill over the course of training for a “baseline” experiment together with a “synchronous” run using only on-policy data (staleness = 0) and restricting each sample to be used at most once (max sample reuse = 1). On the left, the x-axis is wall time. On the right, the x-axis is iterations. Asynchronous training is nearly 3x faster at achieving TrueSkill 150 when measuring by wall time, even though the two runs perform similarly as a function of the number of iterations.

The fact that our algorithms benefit from extremely low sample reuse underlines how sample inefficient they are. Ideally, our training methods could take a small amount of experience and use that to learn a great deal, but currently we cannot even usefully optimize over that experience for more than a couple of gradient steps. Learning to use rollout data more efficiently is one of the major areas for future work in RL research.

This investigation suggests that sample reuse below one can be beneficial. This experiment outperformed all others after around iteration 5,000, including the experiment with sample reuse 1. The improvement over sample reuse 1 is minor compared to the gaps between more severe sample reuses, but it is significant. Intuitively one might expect that using each sample exactly once would be the most optimal, as no data would get wasted and no data would get used twice; collecting more data and then not optimizing over it would not help.

However, the sample reuse is measured as an average rate of data production to consumption (Equation 13). Because the optimizers sample each batch randomly from the buffer, sample reuse 1 just means that on *average* each sample is used once, but in fact many samples are used twice, and some not used at all. For this reason producing twice as much data as we can consume still reduces the number of samples which get selected multiple times. Of course the magnitude of improvement is relatively small and the cost (doubling the number of rollout workers and forward pass GPUs) is significant. Doubling the number of rollout workers may also decrease correlation across samples; using two adjacent samples from the same game (when very little has changed between them) may have similar drawbacks to using the same sample twice.

N Self-play

OpenAI Five is trained without any human gameplay data through a self-improvement process named *self-play*. This technique was successfully used in prior work to obtain super human performance in a variety of multiplayer games including Backgammon, Go, Chess, Hex, StarCraft 2, Poker [1, 4, 7, 37–39]. In self-play training, we continually pit the current best version of an agent against itself or older versions, and optimize for new strategies that can defeat these past and present opponents.

In training OpenAI Five 80% of the games are played against the latest set of parameters, and 20% play against past versions. We play occasionally against past parameter versions in order to obtain more robust strategies and avoid *strategy collapse* in which the agent forgets how to play against a wide variety of opponents because it only requires a narrow set of strategies to defeat its immediate past version (see Balduzzi *et al.* [60] for a discussion of cyclic strategies in games with simultaneous-turns and/or imperfect information).

OpenAI Five uses a dynamic sampling system in which each past opponent $i = 1..N$ is given a *quality* score q_i . Opponent agents are sampled according to a softmax distribution; agent i is chosen with probability p_i proportional to e^{q_i} . Every 10 iterations we add the current agent to past opponent pool and initialize its quality score to the maximum of the existing qualities. After each rollout game is completed, if the past opponent defeats the current agent, no update is applied. If the current agent defeats a past opponent, an update is applied proportional to a learning rate constant η (which we fix at 0.01):

$$q_i \leftarrow q_i - \frac{\eta}{N p_i} \tag{14}$$

In Figure 27 we see the opponent distribution at several points in early training. The spread of the distribution gives a good picture of how quickly the agent is improving: when the agent is improving rapidly, then older opponents are worthless to play against and have very low scores; when progress is slower the agent plays against a wide variety of past opponents.

O Exploration

Exploration is a well-known and well-researched problem in the context of reinforcement learning. We encourage exploration in two different ways: by shaping the loss (entropy and team spirit) and by randomizing the training environment.

O.1 Loss function

Per [14], we use entropy bonus to encourage exploration. This bonus is added to the PPO loss function in the form of $cS[\pi_\theta](s_t)$, where c is a hyperparameter referred to as entropy coefficient. In initial stages of training a long-running experiment like OpenAI Five or Rerun we set it to an initial value and lower it during training. Similarly to [14], [16], or [61], we find that using entropy bonus prevents premature convergence to suboptimal policy. In Figure 28, we see that entropy bonus of 0.01 (our default) performs best. We also find that setting it to 0 in early training, while not optimal, does not completely prevent learning.

As discussed in Appendix G, we introduced a hyperparameter *team spirit* to control whether agents optimize for their individual reward or the shared reward of the team. Early training and speedup curves for team spirit can be seen in Figure 29. We see evidence that early in training,

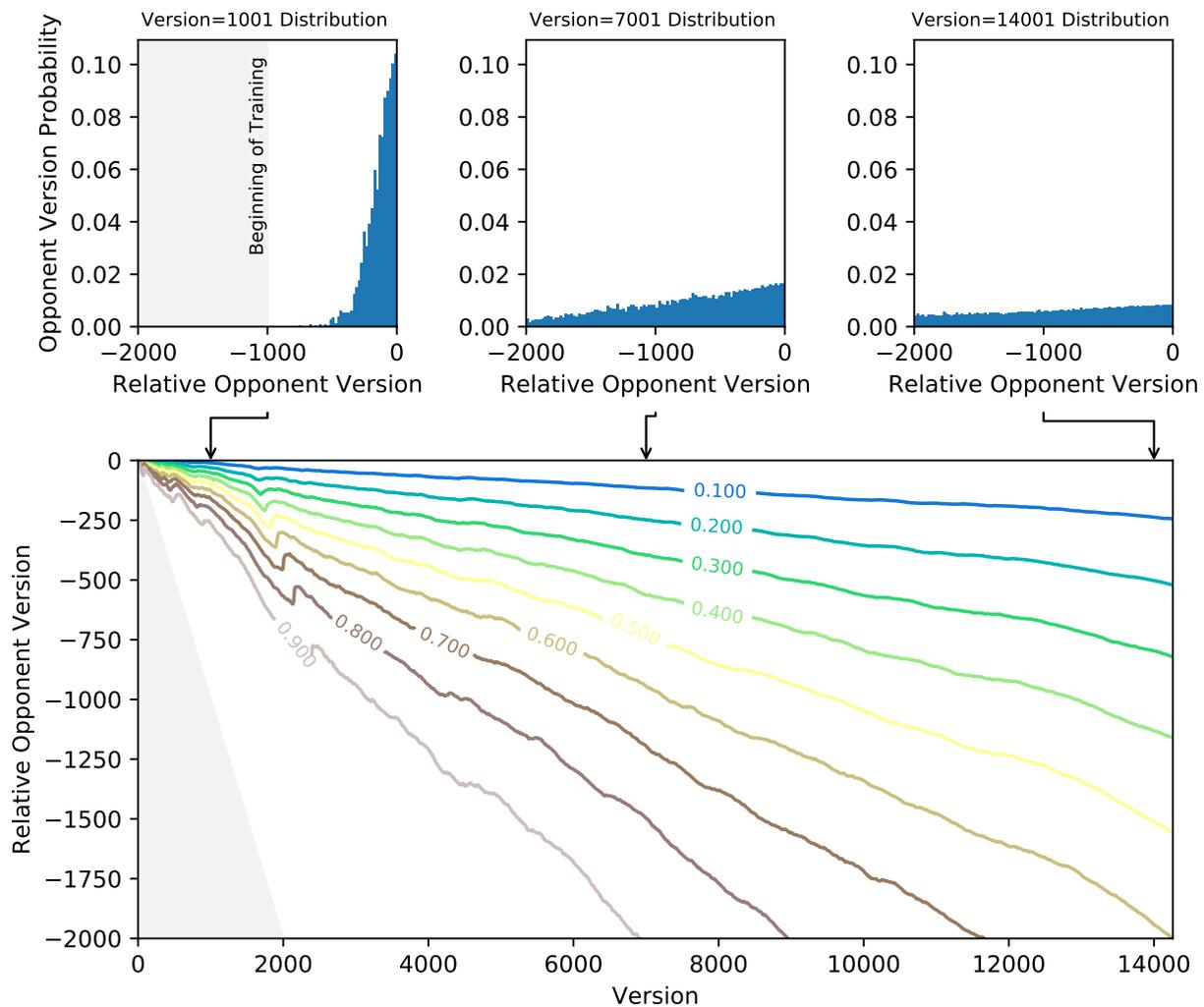


Figure 27: **Opponent Manager Distribution over past versions.** As the performance of the agent improves, the distribution over past versions changes to find stronger contenders. The slope in the distribution reflects how fast the current agent is outpacing previous versions: a slow falloff indicates that the agent is still challenged by far older versions, while a steep falloff is evidence that counter-strategies have been found that eliminate past agents. In later versions the opponent distribution includes many more past versions, suggesting that after a warmup period, skill progression slows.

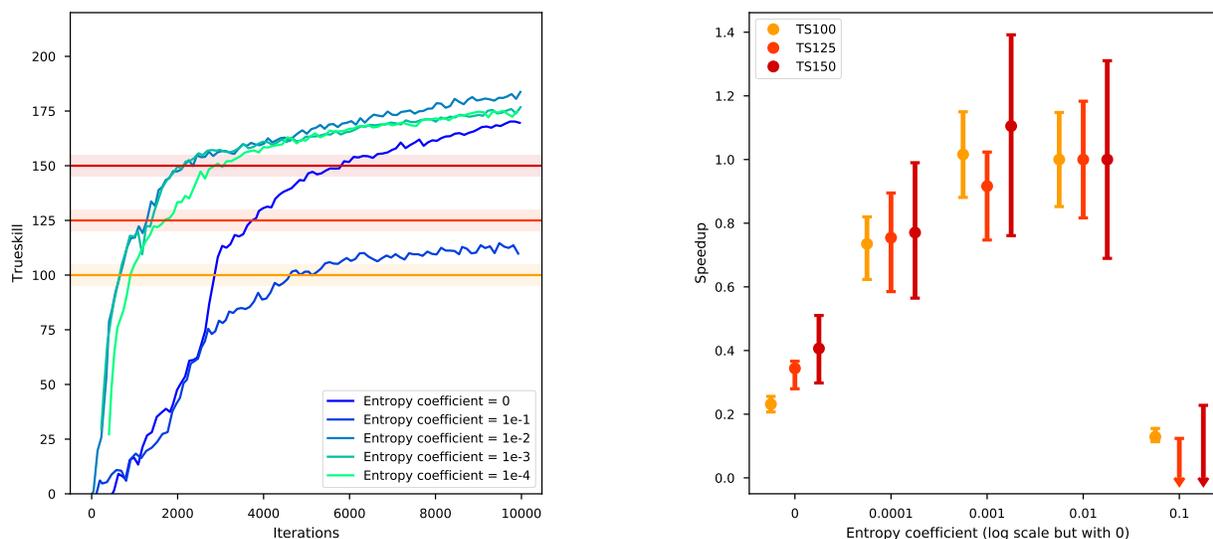


Figure 28: **Entropy in early training:** TrueSkill and speedup with varied entropy coefficients. Lower entropy performs worse because the model has a harder time exploring; higher entropy performs much worse because the actions are too random.

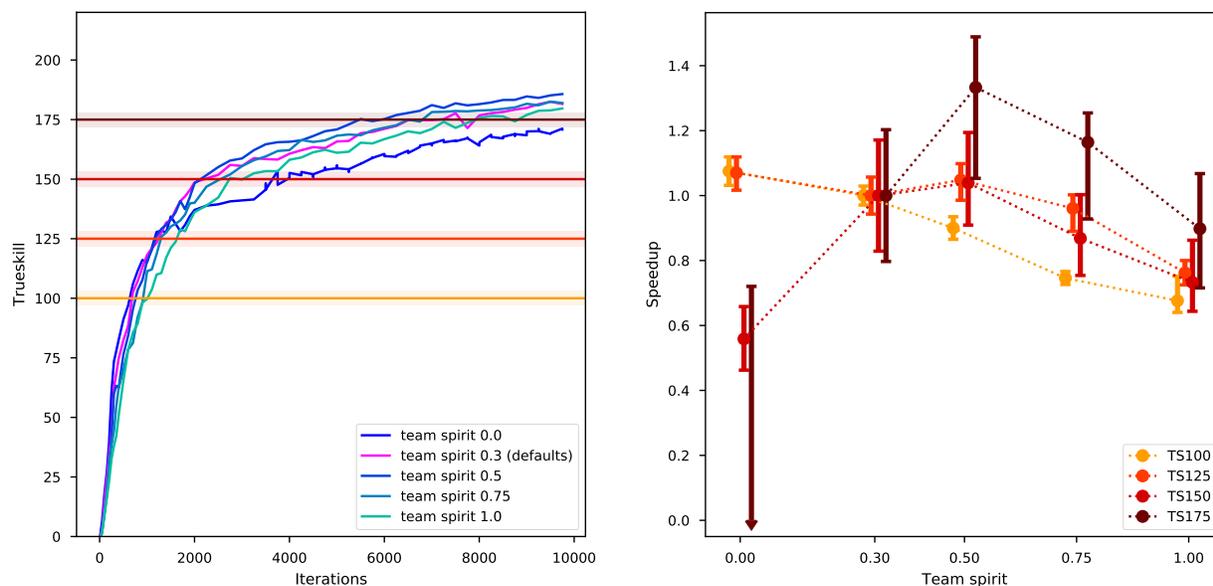


Figure 29: **Team Spirit in early training:** Very early in training (TrueSkill <125) the run with team spirit 0 does best; this can be seen by the speedup for lower TrueSkill being highest at team spirit 0. The maximum speedup quickly moves to 0.5 in the medium TrueSkill regime (150 and 175).

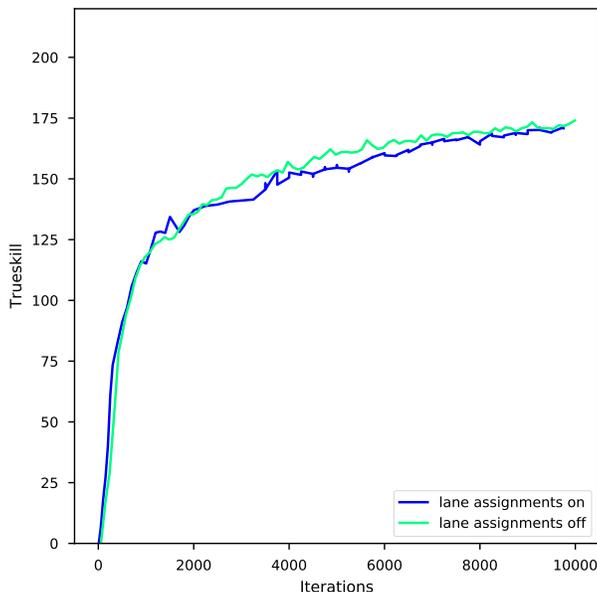


Figure 30: **Lane Assignments:** “Lane assignments” randomization on vs. off. In this ablation we see that this randomization actually provided little benefit.

lower team spirits do better. At the very start team spirit 0 is the best, quickly overtaken by team spirit 0.3 and 0.5. We hypothesize that later in training team spirit 1.0 will be best, as it is optimizing the actual reward signal of interest.

O.2 Environment Randomization

We further encouraged exploration through randomization of the environment, with three simultaneous goals:

1. If a long and very specific series of actions is necessary to be taken by the agent in order to randomly stumble on a reward, and any deviation from that sequence will result in negative advantage, then the longer this series, the less likely is agent to explore this skill thoroughly and learn to use it when necessary.
2. If an environment is highly repetitive, then the agent is more likely to find and stay in a local minimum.
3. In order to be robust to various strategies humans employ, our agents must have encountered a wide variety of situations in training. This parallels the success of domain randomization in transferring policies from simulation to real-world robotics[5].

We randomize many parts of the environment:

- **Initial State:** In our rollout games, heroes start with random perturbations around the default starting level, experience, and gold, armor, movement speed, health regeneration, mana regeneration, magic resistance, strength, intellect, and agility.

- **Lane Assignments:** From a strategic perspective it makes sense for heroes to act in certain area of the map more than the others. Most inter-team skirmishes happen on *lanes* (3 distinct paths that connect opposing bases). At a certain stage of our work, we noticed that our agents developed a preference to stick together as a group of 5 on a single lane, and fighting any opponent coming their way. This represents a large local minimum, with higher short-term reward but lower long-term one as the resources from the other lanes are lost. After that we introduced *lane assignments*, which randomly assigned each hero to a subset of lanes, and penalized them with negative reward for leaving those lanes. However, the ablation study in Figure 30 indicates that this may not have been necessary in the end.
- **Roshan Health:** Roshan is a powerful neutral creature, that sits in a specific location on the map and awaits challengers. Early in training our agents were no match for it; later on, they would already have internalized the lesson never to approach this creature. In order to make this task easier to learn, we randomize Roshan’s health between zero and the full value, making it easier (sometimes much easier) to kill.
- **Hero Lineup:** In each training game, we randomly sample teams from the hero pool. While the hero randomization is necessary for robustness in evaluations against human players (which may use any hero teams), we hypothesize that it may serve as additional exploration encouragement, varying the game and preventing premature convergence. In Appendix P, we see that training with additional heroes causes only a modest slowdown to training despite the extra heroes having new abilities and strategies which interact in complex ways.
- **Item Selection:** Our item selection is scripted: in an evaluation game, our agents always buy the same set of items for each specific hero. In training we randomize around that, swapping, adding, or removing some items from the build. This way we expose our agents to enemies playing with and using those alternative items, which makes our agents more robust to games against human players. There are shortcomings to this method, e.g. a team with randomly picked items is likely to perform worse, as our standard build is carefully crafted. In the end our agent was able to perform well against humans who choose a wide variety of items.

P Hero Pool Size

One of the primary limitations of our agent is its inability to play all the heroes in the game. We compared the progress in early training from training with various numbers of heroes. In all cases, each training game is played using an independent random sampling of five heroes from the pool for each team. To ensure a fair comparison across the runs, evaluation games are played using only the smallest set of heroes. Because the test environment uses only five heroes, the runs which train with fewer heroes are training closer to the test distribution, and thus can be expected to perform better; the question is how much better?

In Figure 31, we see that training with more heroes causes only a modest slowdown. Training with 80 heroes has a speedup factor of approximately 0.8, meaning early training runs 20% slower than with the base 17 heroes. From this we hypothesize that an agent trained on the larger set of heroes using the full resources of compute of Rerun would attain a similar high level of skill with approximately 20% more training time. Of course this experiment only compares the very early stages of training; it could be that the speedup factor becomes worse later in training.

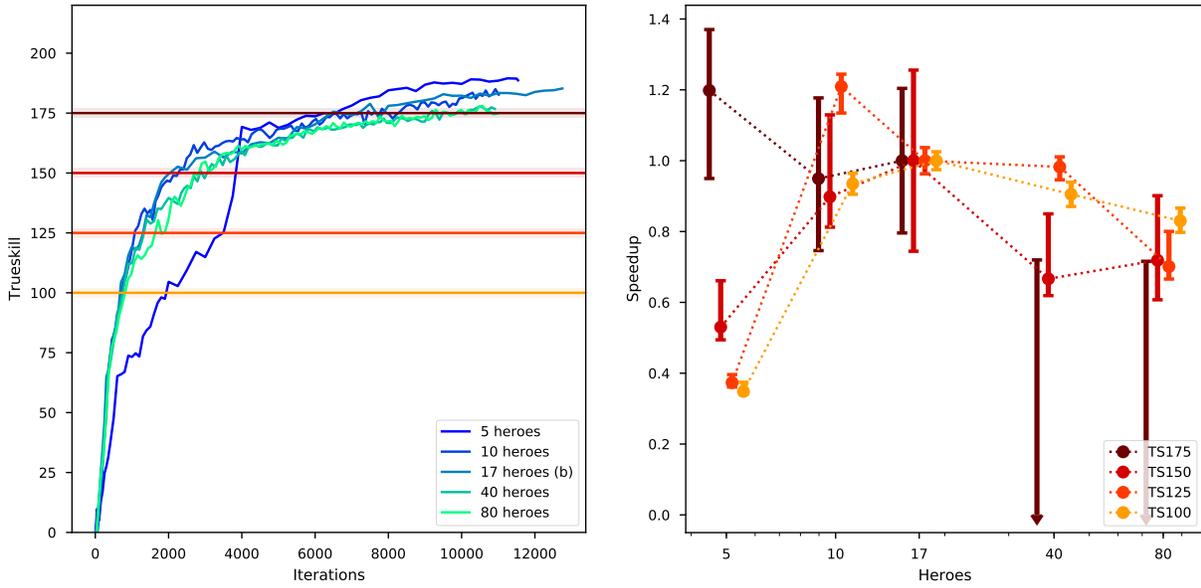


Figure 31: **Effect of hero pool size on training speed:** TrueSkill over the course of training (see Appendix J) and speedup measured by the rate to attain different TrueSkill thresholds (computed using Equation 2) granted by varying the size of the hero pool. Additional heroes slows down early training only slightly. The severe underperformance of the 5-hero run for the first 4k versions was not investigated in detail. It is likely not due to the hero count but rather some instability in that particular training run.



Figure 32: **Learning Rate during The International:** This is what happens when humans under time pressure choose hyperparameters. We believe that in the future automated systems should optimize these hyperparameters instead. After this event, our team began to internally refer to the act of frantically searching over hyperparameters as “designing skyscrapers.”

Q Bloopers

Q.1 Manually Tuned Hyperparameters

Leading into The International competition in August 2018, we already a very good agent, but we felt it was likely not yet as good as the very best humans (as indeed turned out to be the case; we lost both games at that event). In the final few days before the games, we sought to explore high-variance options which had a chance of offering a surprising improvement. In the end, however, we ultimately believe that human intuition, especially under time pressure, is not the best way to set hyperparameters. See Figure 32 for the history of our learning rate parameter during those few days.

Q.2 Zero Team Spirit Embedding

One of our team members stumbled upon a very strange phenomenon while debugging a failed surgery. It turned out that replacing a certain set of 128 learned parameters in the model with zero increased the model’s performance significantly (about 55% winrate after versus before). We believe that the optimizers were unable to find this direction for improvement because although the win rate was higher, the shaped reward (see Table 6) was approximately the same. A random perturbation to the parameters should have overwhelming probability of making things worse rather than better. We do not know why zero would be a special value for these parameters.

These parameters were certainly an unusual piece of the model. In the early stages of applying team spirit (see Appendix G), we attempted to randomize the team spirit parameter in each game. We had a fixed list of four possible team spirit values; each rollout game one was chosen at random. The agent was allowed to observe the current team spirit, via an embedding table with four entries. We hoped this might encourage exploring games of different styles, some very selfless games and some very selfish games.

After training in this way for only a short period, we decided this randomization was not helping, and turned it off. Because our surgery methods do not allow for removing parameters easily, we simply set the team spirit observation to always use a fixed entry in the embedding table. In this way we arrived at a situation where the vector of “global” observations g consisted of the real observations

g_r , concatenated with 128 dimensions from this fixed embedding E ; these extra dimensions were learned parameters which did not depend on the observations state:

$$g = [g_r, E] \tag{15}$$

Because this vector is consumed by a fully connected layer $Wg + B$, these extra parameters do not affect the space of functions representable by the neural network. They are exactly equivalent to not including E in the global observations and instead using a modified bias vector:

$$B' = W [0, E] + B \tag{16}$$

For this reason we were comfortable leaving this vestigial part of the network in place.

Because it was an embedding, there should be nothing special about 0 in the 128-dimensional space of possible values of E . However we see clear evidence that zero *is* special, because a generic perturbation to the parameters should have a negative effect. Indeed, we tried this explicitly — perturbing these parameters in other random directions — and the effect was always negative except for the particular direction of moving towards zero.

Q.3 Learning path dependency

The initial training of OpenAI Five was done as a single consecutive experiment over multiple months. During that time new items, observations, heroes, and neural network components were added. The order of introduction of these changes was a priori not critical to learning, however when reproducing this result in Rerun with all the final observations, heroes, and items included, we found that one item — Divine Rapier — could cause the agents to enter a negative feedback loop that reduced their skill versus reference opponents. As Rapier began to be used, we noted a decrease in episodic reward and TrueSkill. When repeating this experiment with Rapiers banned from the items available for purchase, TrueSkill continues to improve.

We hypothesize that this effect was not observed during our initial long-lasting training because Rapier was only added after the team spirit hyperparameter was raised to 1 (see Appendix G). Rapier is a unique item which does not stay in your inventory when you die, but instead falls to the ground and can be picked up by enemies or allies. Because of this ability to transfer a high-value item, it is possible that the reward collected in a game increases in variance, thereby preventing OpenAI Five from learning a reliable value function.