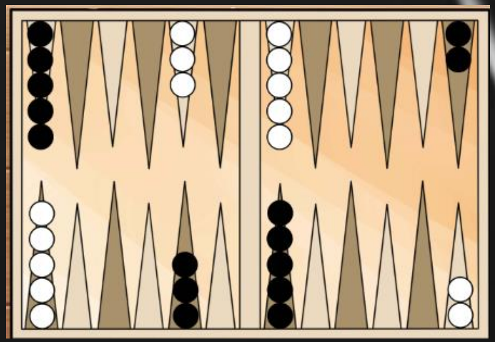
Deep reinforcement learning

Kacper Wnęk, Paweł Świderski, Mateusz Kubita

Agenda

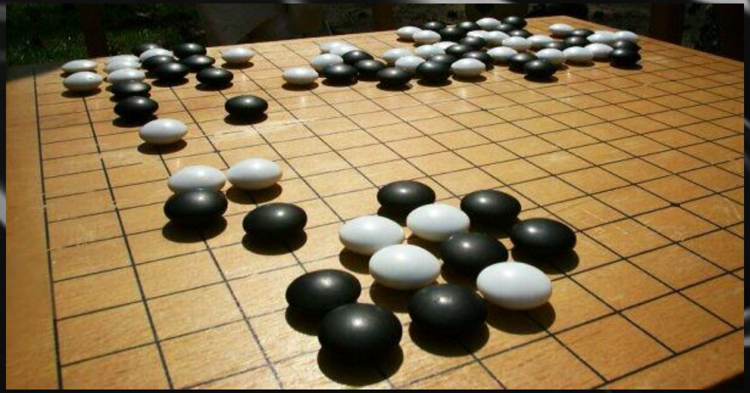
1. **Wprowadzenie**
2. **OpenAI Five: Triumf Sztucznej Inteligencji**
3. **Wyzwania w Dota 2**
4. **Narzędzia Ciągłego Treningu**
5. **Cos się wymysli**

AI vs human



1992 Backgammon

1997 Chess



2016 Go


DOTA 2

Challenges

- **Long time horizons.**
- **Partially-observed state.**
- **High-dimensional action and observation spaces.**


OpenAIFive

Limitations

- **Subset of 17 heroes**
- **No support for items which allow a player to temporarily control multiple units at the same time**

Playing Dota 2 using AI

- Humans interact with the Dota 2 game using three things: keyboard, mose and computer monitor
- The Dota 2 engine runs at 30 frames per second, hence OpenAI Five only acts on every 4th frame. It is called timestamp
- OpenAI collects ~ 16 000 total values each time step


e. Enemy Creep
f. Enemy Heroes
g. Allied Creeps
k. Fog of War
l. Allied Tower
h. Modifiers
i. Items
j. Abilities

Observation space

Global data	22
time since game started	1
is it day or night? time to next day/night change	2
time to next spawn: creep, neutral, bounty, runes	4
time since seen enemy courier is that > 40 seconds?[a]	2
min&max time to Rosh spawn	2
Roshan's current max hp	1
is Roshan definitely alive?	1
is Roshan definitely dead?	1
Next Roshan drops cheese?	1
Next Roshan drops refresher?	1

Per-hero add'l (10 heroes)	25
is currently alive?	1
number of deaths	1
hero currently in sight? time since this hero last seen	2
hero currently teleporting? if so, target coordinates (x, y) time they've been channeling	4
respawn time	1
current gold (allies only)	1
level	1
mana: max, current, & regen	3
health regen rate	1
magic resistance	1

Per-modifier (10 heroes x 10 modifiers & 179 non-heroes x 2 modifiers)	2
remaining duration	1
stack count	1
modifier name	1

Per-item (10 heroes x 16 items)	13
location one-hot (inventory/backpack/stash)	3
charges	1
is on cooldown? cooldown time	2
is disabled by recent swap?	

Reward weights

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.

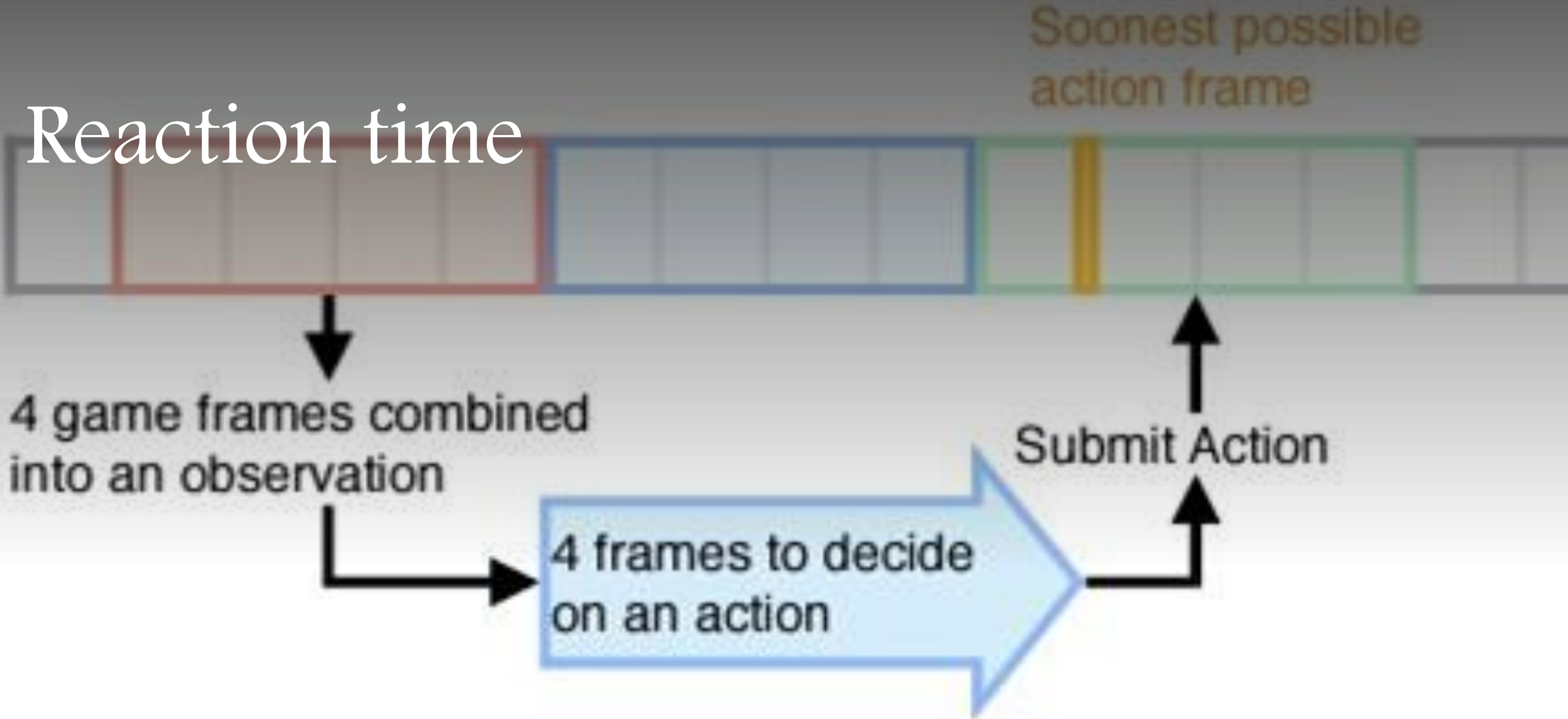
Game time weighting

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

Team spirit

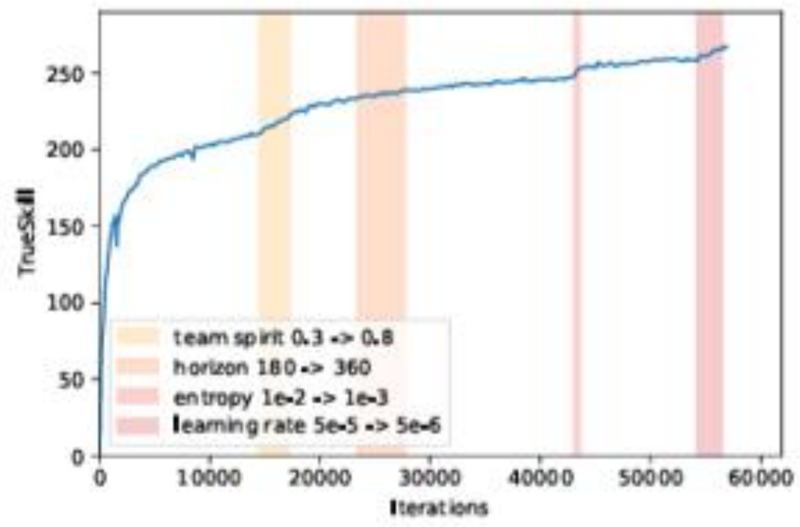
$$r_i = (1 - \tau)\rho_i + \tau\overline{\rho}$$

Reaction time



Hyperparameter changes over time and optimizing the policy

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



Surgery methodology

Can u beat it?

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.

Eperiments and Evaluation

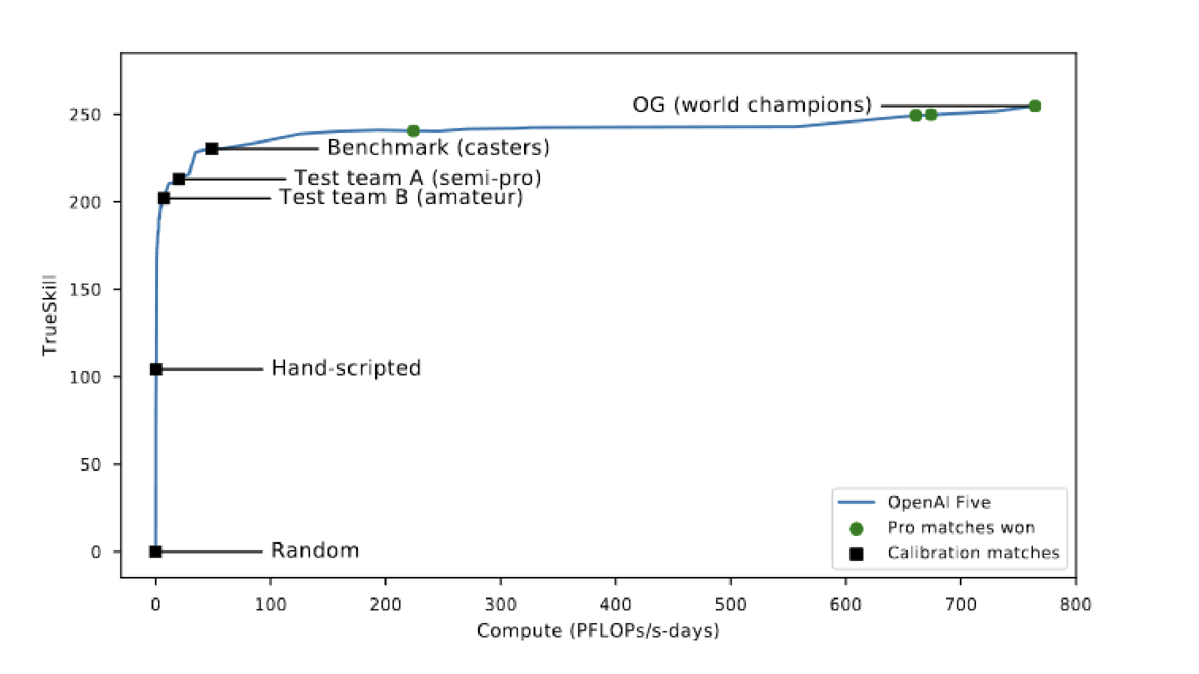
- AI uczyło się 10 miesięcy od 30 czerwca 2018 do 22 kwietnia 2019
- Pokonało mistrzów świata oraz 99.4% graczy
- Dużo większa skala niż podobne poprzednie AI jak np. AlphaGo

Human Evaluation

OpenAI Five pokonało mistrzów świata 13 kwietnia 2019

18–21 kwietnia OpenAI Five Arena

TrueSkill



OG (world champions)
Benchmark (casters)
Test team A (semi-pro)
Test team B (amateur)
Hand-scripted
Random
TrueSkill
0
50
100
150
200
250
0
100
200
300
400
500
600
700
800
Compute (PFLOPs/s-days)
OpenAI Five
Pro matches won
Calibration matches

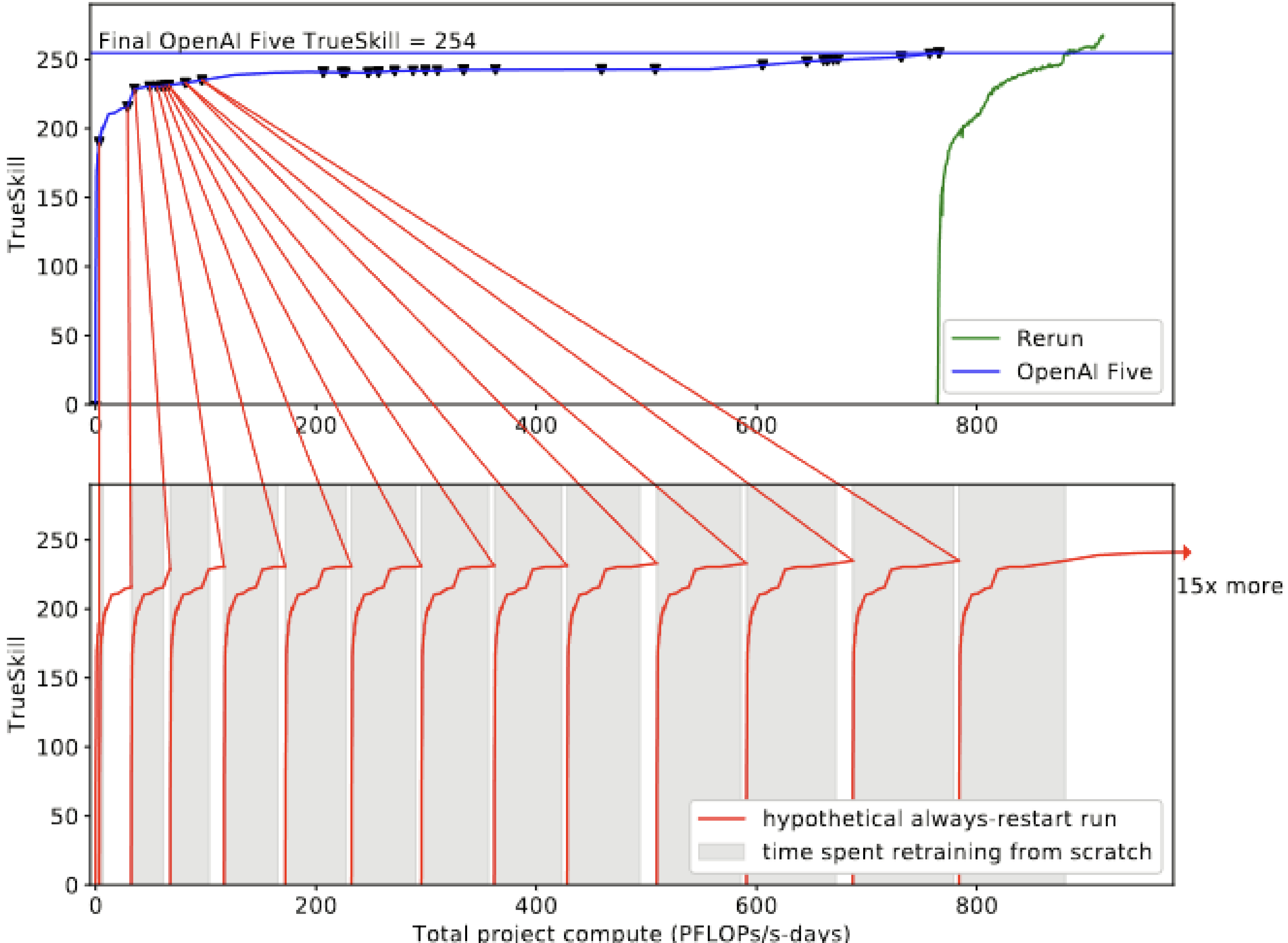
Playstyle

Na początku unikalny, miał swoje wady

Z czasem był bardziej ludzki ale nadal miał unikalne dla siebie zachowania

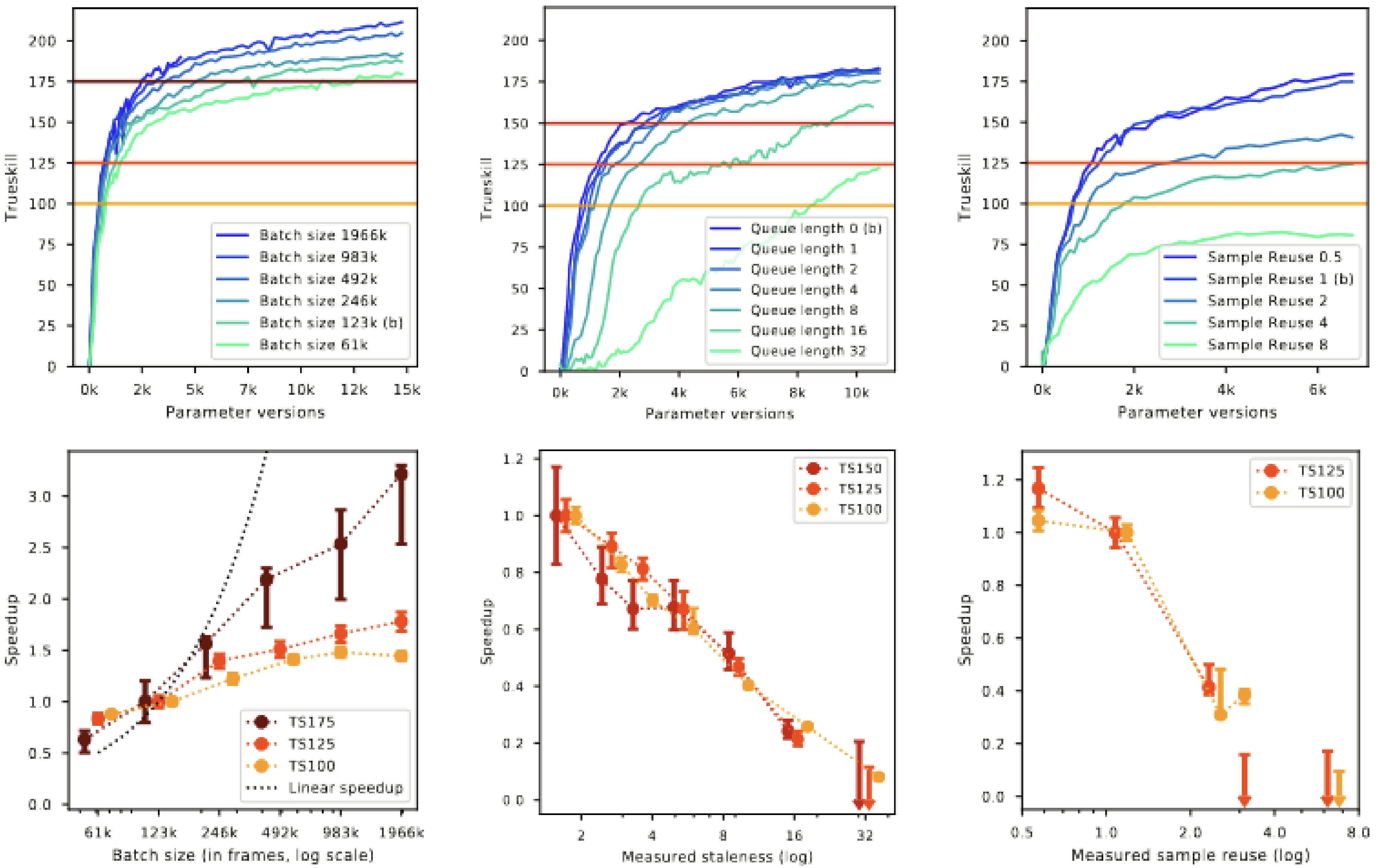
Validating surgery with rerun

- Rerun trwał 2 miesiące
- Jeżeli zamiast surgery wykonywany był restart uczenia, AI uczyło by się 40 zamiast 10 miesięcy
- Rerun miał 98% win rate z OpenAI Fice
- Surgery jest potencjalnie do poprawienia, gdyż rerun osiąga wyższy TS


Final OpenAI Five TrueSkill = 254
TrueSkill
Rerun
OpenAI Five
15x more
hypothetical always-restart run
time spent retraining from scratch
Total project compute (PFLOPs/s-days)

Batch size

- Zwiększenie liczby GPU oraz liczby rollout machines
- Pożądany speedup spowodowany zwiększaniem batch size jest liniowy
- Liniowy speedup nie został osiągnięty, jednak był on znaczący



Trueskill
Parameter versions
Batch size 1966k
Batch size 983k
Batch size 492k
Batch size 246k
Batch size 123k (b)
Batch size 61k
Queue length 0 (b)
Queue length 1
Queue length 2
Queue length 4
Queue length 8
Queue length 16
Queue length 32
Sample Reuse 0.5
Sample Reuse 1 (b)
Sample Reuse 2
Sample Reuse 4
Sample Reuse 8
Speedup
Batch size (in frames, log scale)
TS175
TS125
TS100
Linear speedup
Measured staleness (log)
TS150
TS125
TS100
Measured sample reuse (log)
TS125
TS100

Data Quality

Około 2h trwania jednej gry powoduje problemy podczas uczenia

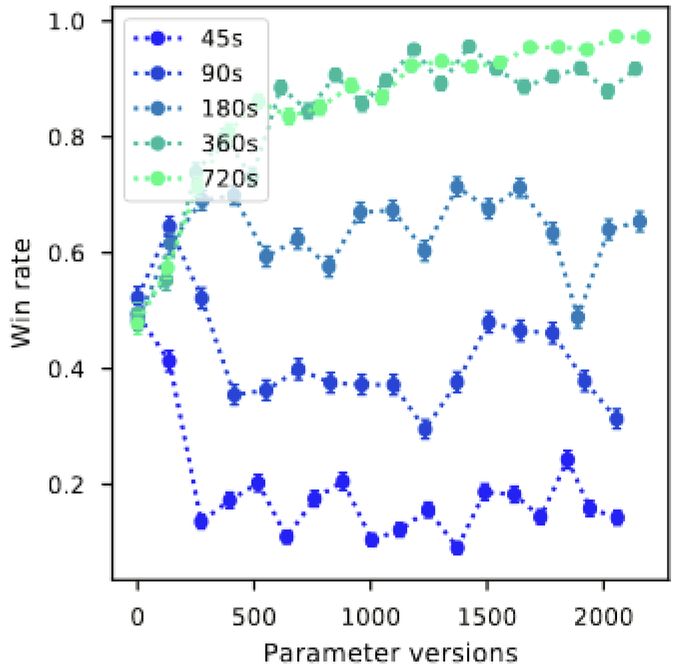
Twórcy użyli podejścia asynchronicznego w procesie uczenia

Pojawiły się problemy staleness i sample reuse

Long term credit assignment

- Dota 2 ma bardzo długi czas zależności między ruchami
- Time horizon

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



Podsumowanie

- OpenAI Five osiągnęło nadludzki poziom w grze Dota2
- Kluczowym składnikiem było zwiększenie batch size'u oraz czasu uczenia z użyciem surgery
- Możliwe, że rezultaty mogą być przełożone na inne gry
- Wraz ze zwiększaniem się złożoności problemów i środowisk skalowanie będzie jeszcze ważniejsze

Bibliografia

- Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model", ArXiv preprint arXiv:1912.06680, 2019.
- Hugging Face: 2. Hugging Face, "Introduction to Deep Reinforcement Learning", Hugging Face Blog, [dostęp: 13 kwietnia 2024], https://huggingface.co/blog/deep-rl-intro.
- Andrew Trask, "Deep Learning from Scratch: Building with Python from First Principles", O'Reilly Media, 2019.

Dziękujemy za uwagę