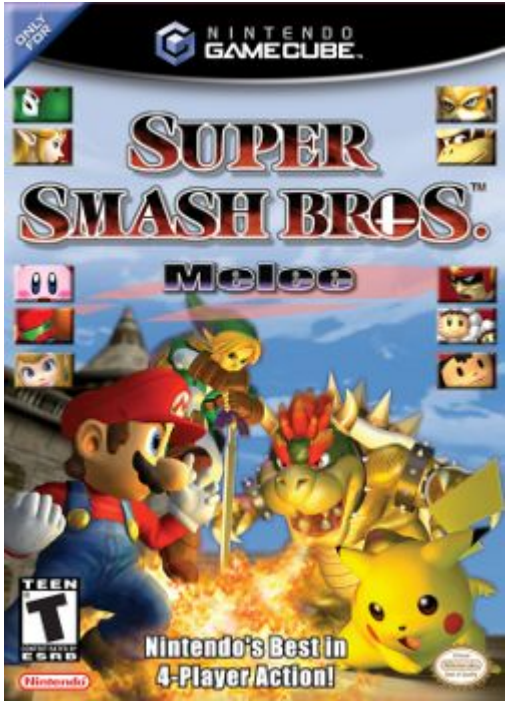
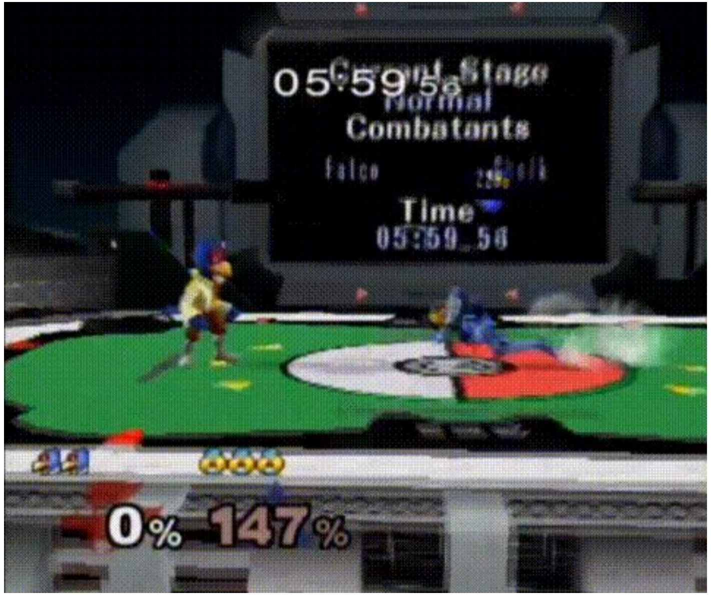
Tackling Super Smash Bros. Melee with Deep RL

Vlad Firoiu


ONLY FOR
NINTENDO GAMECUBE
SUPER SMASH BROS.™
Melee
TEEN
T
ESRB
Nintendo's Best in 4-Player Action!
Nintendo

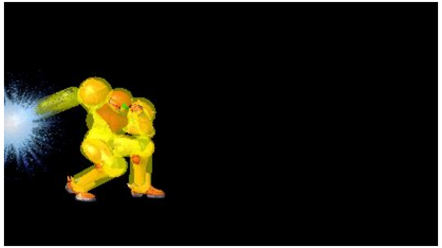
NINTENDO GAMECUBE
NINTENDO GAMECUBE
Nintendo

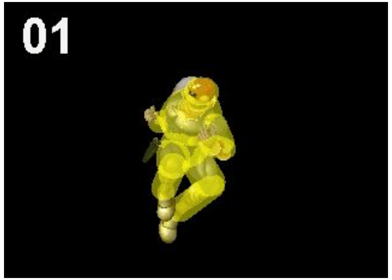


A series of attacks ending in a KO.

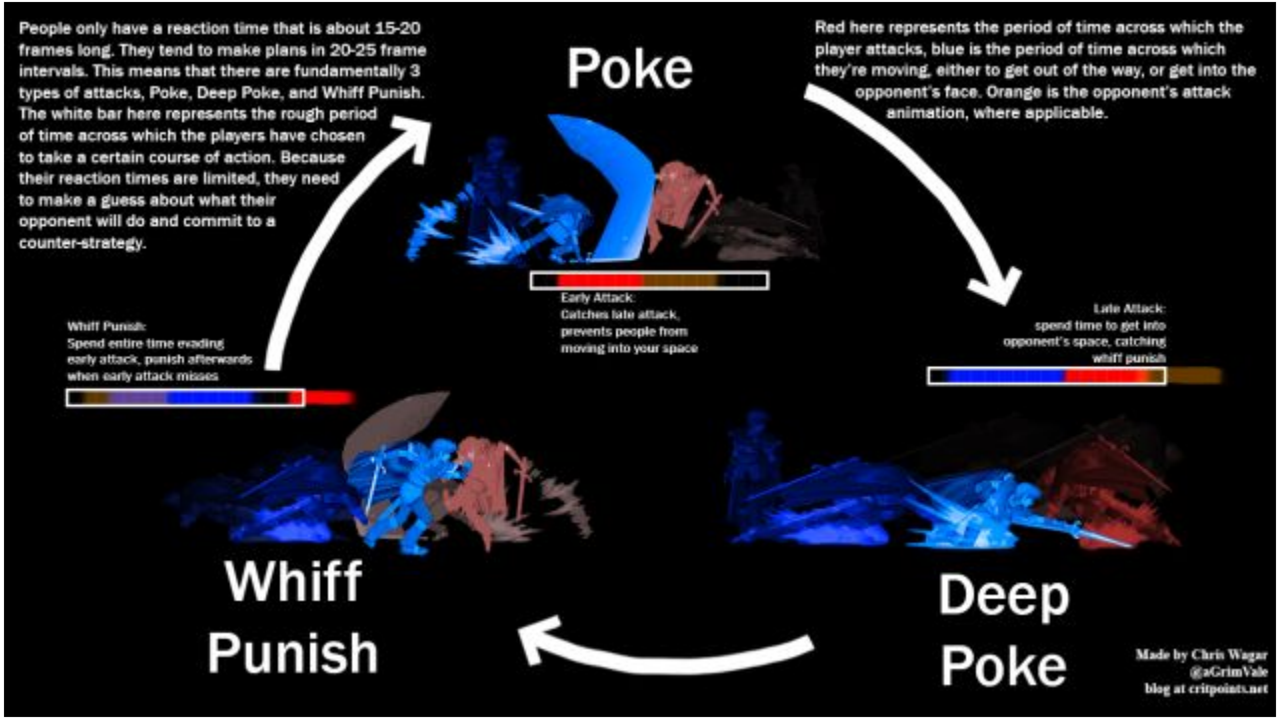
Top-level gameplay. Not sped up!

“Yomi” in Melee





“Yomi” in Melee



People only have a reaction time that is about 15-20 frames long. They tend to make plans in 20-25 frame intervals. This means that there are fundamentally 3 types of attacks, Poke, Deep Poke, and Whiff Punish. The white bar here represents the rough period of time across which the players have chosen to take a certain course of action. Because their reaction times are limited, they need to make a guess about what their opponent will do and commit to a counter-strategy.

Red here represents the period of time across which the player attacks, blue is the period of time across which they're moving, either to get out of the way, or get into the opponent's face. Orange is the opponent's attack animation, where applicable.

Early Attack:
Catches late attack,
prevents people from
moving into your space

Late Attack:
spend time to get into
opponent's space, catching
whiff punish

Whiff Punish:
Spend entire time evading
early attack, punish afterwards
when early attack misses

Made by Chris Wagar
@aGrimVale
blog at critpoints.net

The RL Environment

- Environment simulated with the Dolphin emulator
  - 1-2x real time on server hardware (no graphics)
- Game state is read from RAM on each frame
  - Player positions, velocities, facing direction, etc
  - 382 (!) action states (running, jumping, attacking)
  - mostly observable
- Action space
  - Full controller has 2 analog control sticks, 7 buttons, 2 triggers.
  - Compressed down to ~50 discrete actions.
- Reward structure
  - +/- 1 for kills/deaths
  - +/- 0.01 for damage dealt/taken

Methods

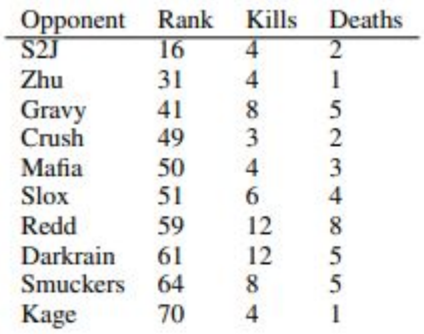
- RL: Importance-Weighted Advantage Actor-Critic (IMPALA)
    - 50-200 environments per experiment.
    - Early versions had no importance weighting.
- DL: small MLP
    - Initially frame stacking.
    - Later added a GRU.
- Symmetric self-play
    - Against exact same parameters → double throughput.
    - Against mixture of old checkpoints.
        - Indicates rate of improvement over time.
    - Modern solution is AlphaStar League.

Results

After 1 day: garbage. Left the experiment running and forgot about it.

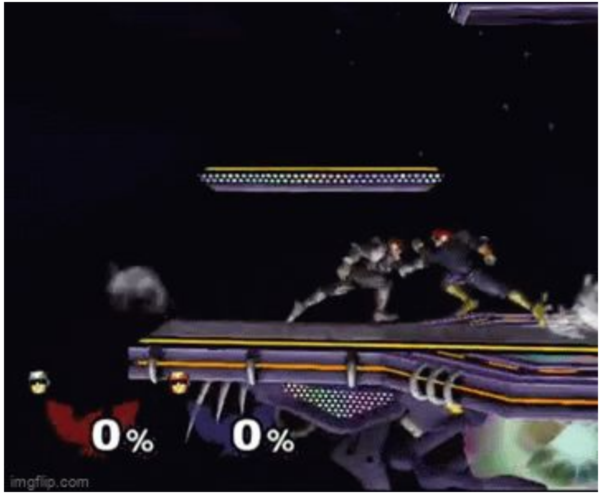
After 1 week… very strong, better than me.

After 1 more week, could beat professional players.

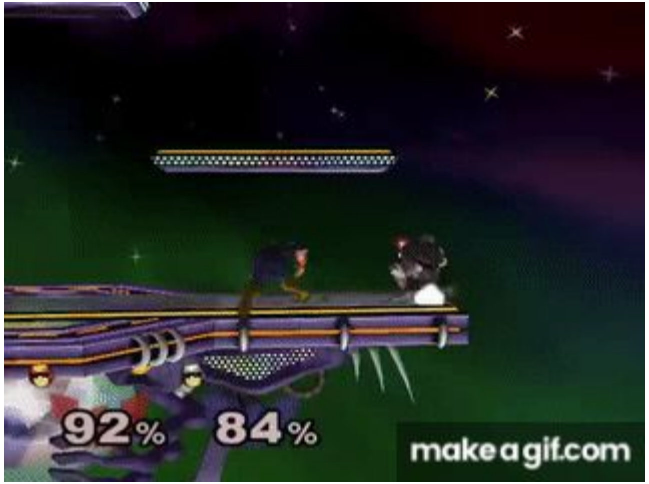


Opponent	Rank	Kills	Deaths
S2J	16	4	2
Zhu	31	4	1
Gravy	41	8	5
Crush	49	3	2
Mafia	50	4	3
Slox	51	6	4
Redd	59	12	8
Darkrain	61	12	5
Smuckers	64	8	5
Kage	70	4	1

Limitations



A simple exploit.

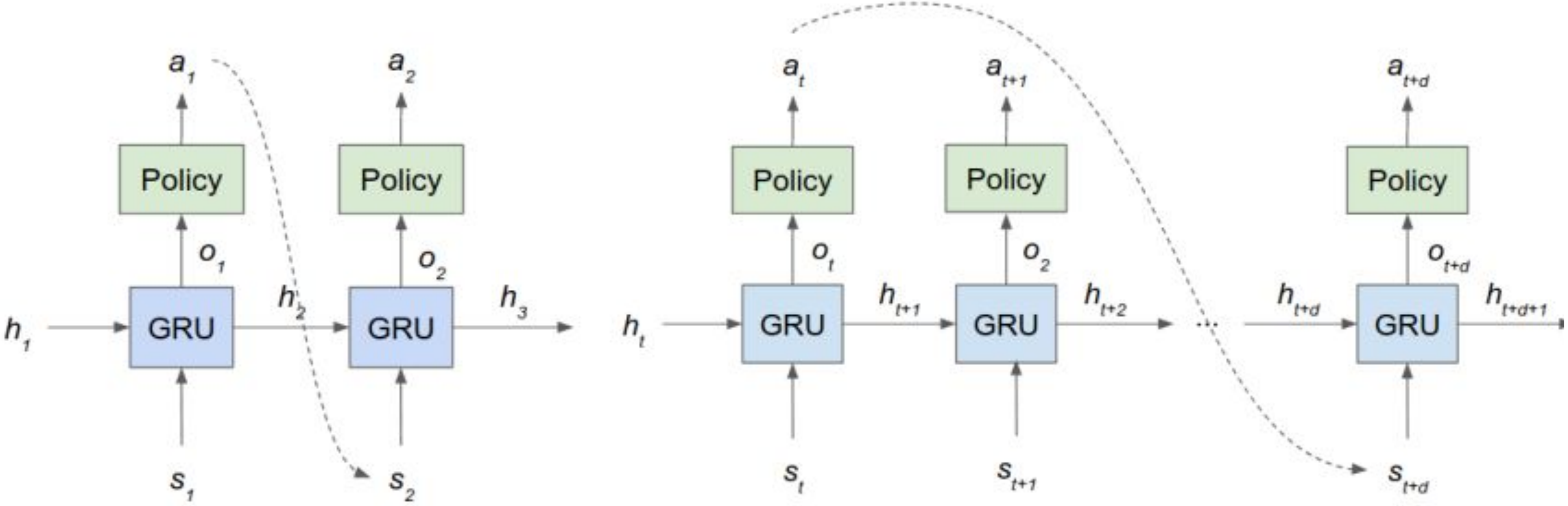


Inhuman speed and reaction time.

Deep RL with Action Delay

- Action delay levels the playing field between human and AI
  - Removes “degenerate” inhuman behavior
- Humans have ~300ms visual reaction time
  - Varies with complexity of task (250-800)
  - Corresponds to 6 agent actions
- Issues for Deep RL
  - Makes credit assignment harder
  - Makes control much harder
  - Correct action depends on heavily on unknown future state

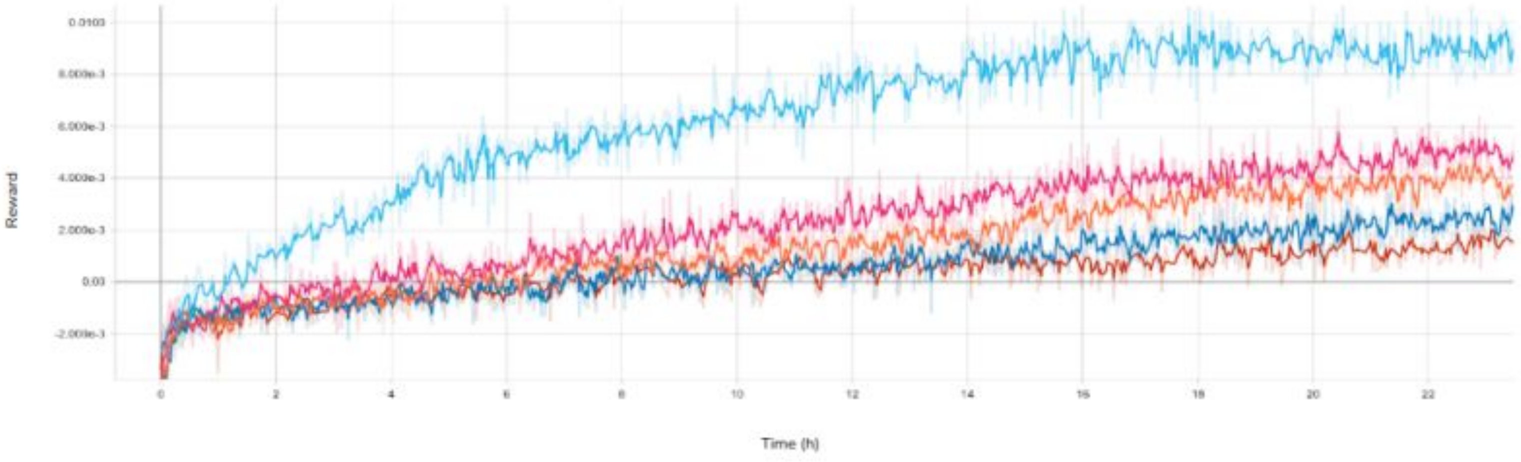
Deep RL with Action Delay



(a) An agent unrolled over time.

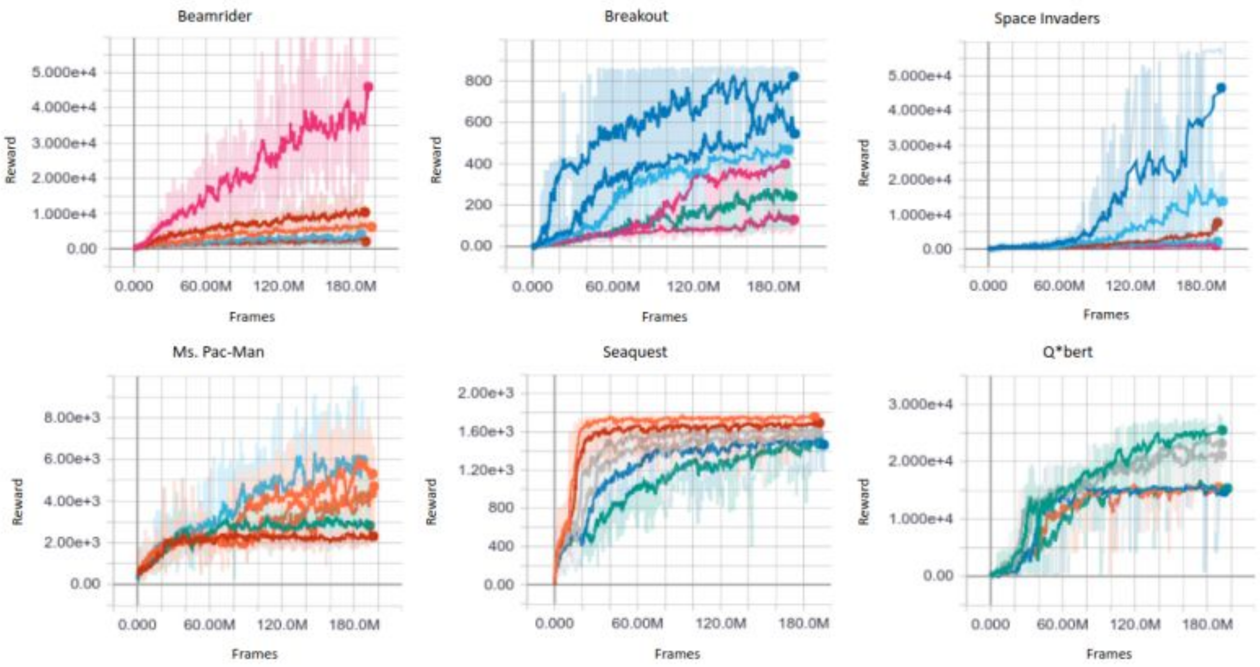
(b) A delayed agent unrolled over time.

The Action Delay Problem: SSBM



Training vs. the in-game AI with delays of 0 (light blue), 1(magenta), 2 (orange), 4 (dark blue) and 5 (brown) agent steps. Each agent step is 50 ms.

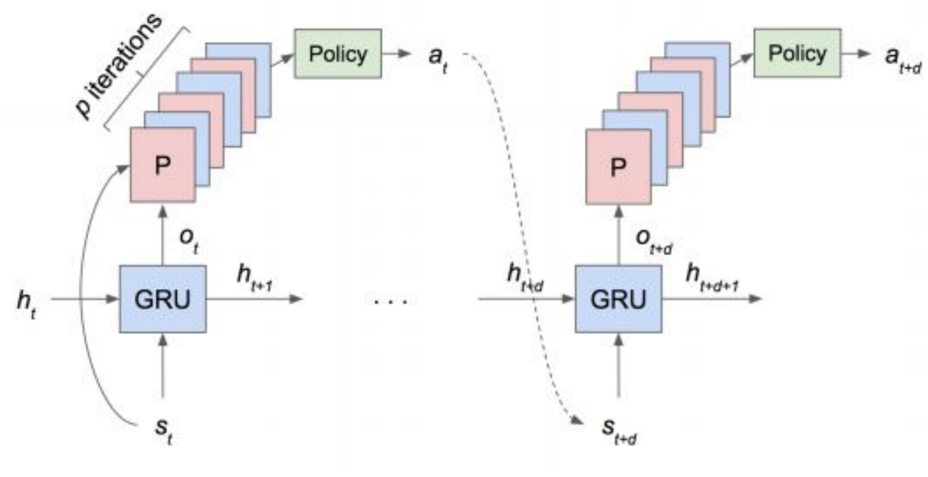
The Action Delay Problem: Atari



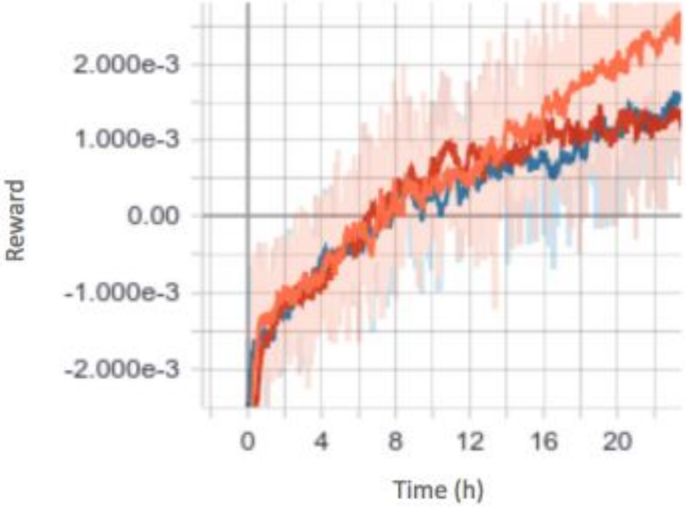
IMPALA trained on Atari with between 0 and 5 steps of delay (up to 333ms).

Solution: “Undoing” Delay

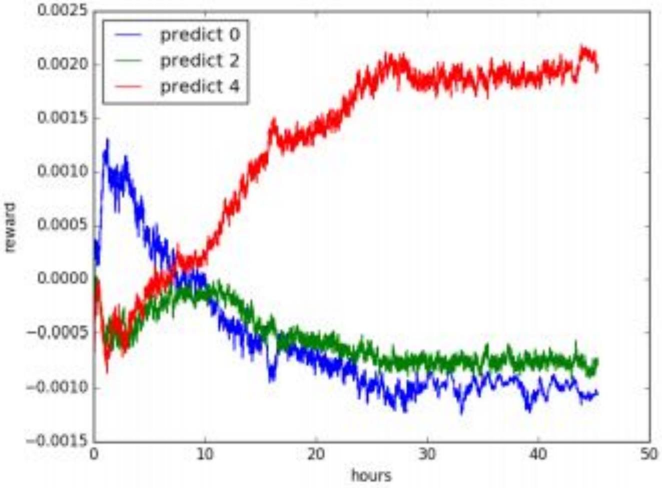
- We perceive moving objects slightly *ahead in space*.
  - See “Flash-Lag Effect”, an optical illusion.
- Use a learned environment model to do the same.
  - Gives the policy a better estimate of the future state.



Results of Undoing Delay



Delay 4 against in-game AI, predicting 0 (blue), 2 (red) and 4 (orange) frames into the future.



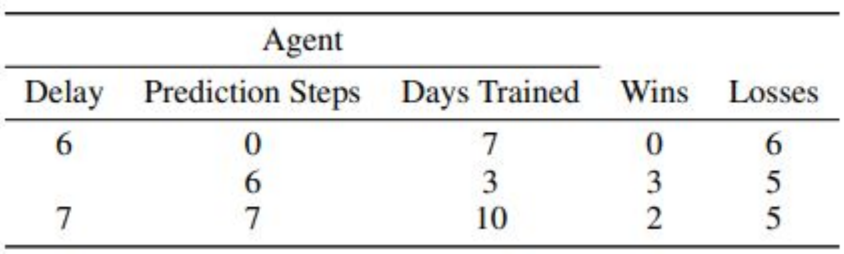
Three delay=4 agents co-training for two days.

Results vs. Human Opponents

Good, but not superhuman.

More “human-like” than before, but still relies on very precise reactions.

Later agents went up to 300ms, but fell short of pro level.



Agent				
Delay	Prediction Steps	Days Trained	Wins	Losses
6	0	7	0	6
	6	3	3	5
7	7	10	2	5

Performance against Professor Pro, a top-50 player.
Each win/loss is in a 4-stock match.
For this agent, 7 frames = 233ms.

Remarks & Future Work

- Delay not solved yet
  - Opponent is considered part of environment
  - Could combine MuZero with delay
- Exploration not solved yet
  - RL alone can’t discover some important techniques and strategies
  - OpenAI 5 needed heavy reward shaping
  - AlphaStar used millions of human games
- Imitation learning for Smash Bros
  - Already a dataset of 100K tournament matches
  - Many more games are being recorded in the online “covid” era