Bracket Stamina: Inferring the Intent of Other Agents in a Multiplayer Kelly Coin Flip Game

Eryk Banatt
Johns Hopkins University
Laurel, MD 20723
ebanatt1@jh.edu

Cameron LaFortune
Johns Hopkins University
Laurel, MD 20723
clafort2@jh.edu

Abstract

In multi-agent systems, understanding how artificial agents infer the intentions of others is crucial for effective collaboration and competition. We investigate the decision-making processes of neural networks in the context of the Bracket Stamina game, a novel multiplayer generalization of the Kelly Coin Flip game, with a focus on the concepts of Simulation Theory (ST) and Theory Theory (TT). The game requires players to strategically manage their resources while inferring their opponents' decision-making processes, making it an ideal setting to study social cognition in goal-oriented agents. We design a Deep Q Network (DQN) that learns a favorable strategy for the game, outperforming a random policy even with a lower points budget. To understand the extent to which the DQN can adapt its strategy to diverse decision-making processes, we evaluate its performance against various agent types, including human agents and simple heuristic-based models. Our findings provide valuable insights into the representation and learning of decision-making processes in neural networks, highlighting the potential for developing more adaptive artificial agents in complex, socially-driven environments, and contributing to our understanding of ST and TT in artificial systems. Our code can be found on Github.

1 Introduction

The study of decision-making processes in artificial agents has garnered considerable attention in recent years, particularly as these agents have been increasingly deployed in various applications, ranging from game-playing [18] to natural language understanding [3]. One classic game used to study decision-making and optimal resource allocation is the Kelly Coin Flip game [9], which has been extensively researched due to its simplicity and relevance to real-world problems. In the game, players repeatedly bet on the outcome of a weighted coin flip, aiming to maximize their wealth over time by making smart bets. The game provides an ideal framework for studying optimal betting strategies, risk management, and the rational resource management under uncertainty.

Despite its simplicity, the Kelly Coin Flip game has been a useful tool for understanding various aspects of decision-making in both humans and artificial agents. The game has been employed to study reinforcement learning algorithms and their convergence to optimal strategies [2], as well as the impact of different types of information on decision-making processes. However, the Kelly Coin Flip game primarily focuses on single-agent decision-making and does not directly address the challenge of inferring the intentions of other agents in a multi-agent setting.

To bridge this gap and extend the study of decision-making processes to multi-agent systems, we introduce the **Bracket Stamina** game, a novel multiplayer generalization of the Kelly Coin Flip game. The game not only demands strategic thinking from its players but also requires them to infer their opponents' decision-making processes to succeed. By exploring the underlying "psychology" of

Preprint. Under review.

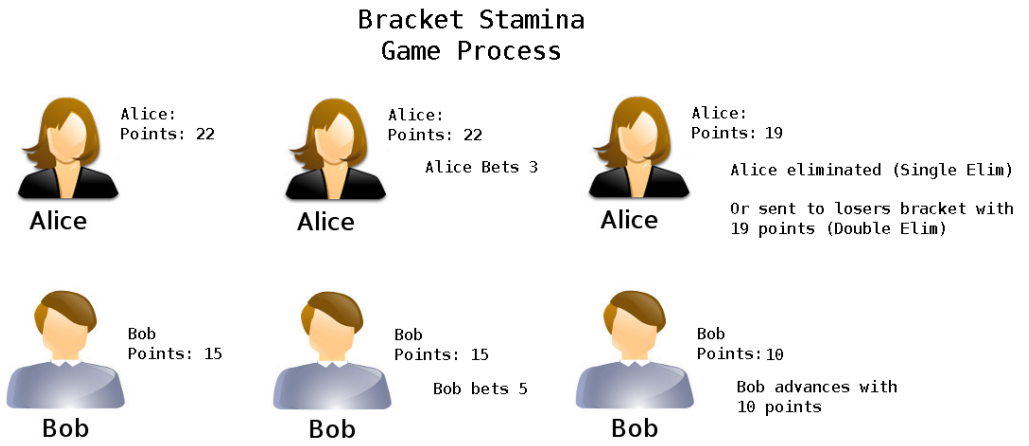

Figure 1: This is a placeholder figure. It will be replaced with a figure depicting bracket stamina

neural networks, we aim to understand the emergent "decision-making processes" of large language models and other artificial agents, which has broad implications in the design and deployment of these systems in various applications.

Our primary contributions are threefold: First, we propose a novel game called "Bracket Stamina" that extends the Kelly Coin Flip game to a multiplayer setting, where agents must not only optimize their bet sizes but also infer the intentions of their opponents in a diverse pool of possible opponents. Second, we develop a Deep Q Network (DQN) [11] that learns a favorable strategy for the Bracket Stamina game, significantly outperforming a random policy even with a much lower points budget. Finally, we evaluate the performance of the DQN when confronted with different agent types, including human agents and simple heuristic-based models, in order to understand the extent to which the DQN can adapt its strategy to diverse opponent decision-making processes.

Through these contributions, we hope to gain insights into the representation and learning of decision-making processes in neural networks, paving the way for the development of more robust and adaptive artificial agents in complex, socially-driven environments.

2 Bracket Stamina Game

The Bracket Stamina game is a multiplayer generalization of the Kelly Coin Flip game, where each player begins with a randomly assigned number of points called "stamina". Players are placed into a seeded, single-elimination bracket based on their number of points. In each round, two players are paired together and have to select a number of their points to spend. If they pick a number that is lower than their opponent's, they are eliminated from the tournament. If they pick a number that is higher than their opponent's, those points are deducted from the player's stamina and they advance to the next round. If the two pick the same number, a winner is randomly selected.

The objective of the game is to be the last player remaining by conserving points and narrowly defeating opponents until the last round. Table 1 presents the probability of victory for agents with random policies as a function of their initial stamina.

3 Related Work

3.1 Iterative Games

Bracket Stamina can more specifically be described as a cross between the Kelly Coin Flip Game [9] and the Iterated Prisoners' Dilemma [16].

In the Kelly Coin Flip Game, a single player is given a weighted coin and a ceiling on the number of possible flips, and can bet any amount of their current wealth on the toss of a coin. The original

Initial Stamina	Probability of Victory
5	0.003
10	0.026
15	0.063
20	0.107
23	0.155
DQN (15)*	0.144

Table 1: Probability of Victory for Random Policies. All other agents start with 15 stamina, the agent of interest starts with the value in the lefthand column. All agents use a random policy except for the final agent, which uses the DQN with 15 stamina.

work for this experiment was primarily focused on an exploration of rational decision-making under a priori known values of uncertainty, and in Haghani and Dewey [9] a modified version of the Kelly Criterion [10] as a heuristic which will hit the ceiling with high likelihood.

The Iterated Prisoners' Dilemma [4] is a multiplayer social game where the classic prisoners' dilemma is played multiple rounds in a row against the same opponent. Unlike in the traditional prisoners' dilemma, where the players converge upon a Nash equilibrium [12] encouraging both players to always defect, in the iterated setting cooperation can outperform always defecting over time. More importantly for our setting, Axelrod [1] showed that strong strategies could be constructed based on the previous decisions made by a player's opponent.

Our setting is one which bridges the gap between these two well-known games. It replaces the a priori known uncertainty value of the Kelly Coin Flip Game with the noisy decisions made by another agent in a social setting. In our setting, the state space is both larger and continuous compared to the iterated prisoners' dilemma, while maintaining a generally iterative structure which allows for strategies to evolve over time. Fundamentally, we describe a setting where, like Axelrod [1], the primary mechanic revolves around predicting your opponent's action while simultaneously extending the state space to allow for richer and more expressive emergent behaviors.

3.2 Artificial Players for Iterative Games

Artificial players for simple iterative games like the Kelly Coin-Flip Game and the Iterated Prisoners' Dilemma have been a prominent component of research into these settings since their original conception. Axelrod [1] pit a number of simple heuristic agents together in a large "tournament" for iterated prisoners' dilemma, in order to evaluate which methods produce the highest average returns. Rapoport et al. [15] further investigates the distribution of "types" of players, the overall hostility of the tournament, and alternative scoring metrics, all of which were shown to have a noticeable effect on the success of different strategies.

Similar to our agent, Sandholm and Crites [17] use Q-learning as a mechanism for finding a strong strategy for the iterated prisoners' dilemma. In that work, cooperation is an explicitly rewarded behavior for their model, in order to avoid an unstable learning environment. This stands in contrast to our setting, a setting seeking to understand that very unstable environment, where responding to an opponent's evolving strategy and changing your own strategy in response may be considered a natural and indeed optimal skill to learn. Likewise, this work has a much clearer, shared reward function that all agents share (number of points accumulated in total), which again differs from our setting where the uncertain and perhaps different incentive structure of the other agent is a critical component to sizing your bet.

Branwen et al. [2] describes a machine learning approach to optimal play in the Kelly Coin Flip Game, where decision trees are used to learn a value function which accounts for the winnings cap and the number of rounds. In that work, they also describe a Generalized Kelly coin-flip game, which randomizes and obscures the weight / winnings cap / number of rounds, turning the game into a partially observable Markov decision process (POMDP), as well as some attempts at using deep reinforcement learning to solve this generalized version of the game.

Some work has additionally been done for achieving superhuman play in the game Diplomacy by combining a large language model [5] [14] with a strategic reasoning module. [6]. While large

language models themselves are outside the scope of this work, they do highlight the need for work on analyzing the internal states of machine learning models. While Meta's CICERO can directly use language to accomplish goals in a social environment, it's unclear what the "internal representation" of the other agents is, if such a representation even exists at all. As language model output grows more human-like, an understanding of the perspective-taking capabilities of these models in their interactions with humans grows to a greater degree of importance.

3.3 Simulation Theory & Theory Theory

Simulation Theory (ST) and Theory Theory (TT) are models for explaining how agents predict the actions and internal states of other agents. With respect to artificial agents in social games, these theories are useful a useful framework for understanding the latent and often difficult to interpret decision-making processes.

Simulation Theory posits that humans understand others by simulating their mental states and decision-making processes within their own minds, essentially putting themselves in the other's position [7]. This approach relies on the agent's own cognitive resources and can be computationally efficient to implement. Theory Theory suggests that humans develop abstract theories about the mental states and decision-making processes of others, independent of their own cognitive processes [8]. This approach can be computationally demanding but may enable a more flexible understanding of diverse opponents.

Our setting seeks to describe a setting where traditional reinforcement learning paradigms will fail unless all other agents follow similar strategies to the agent itself. That is, a game which *requires* theory theory, rather than simulation theory. We aim to highlight a gap in "psychology" between human players and machine players, wherein human players will immediately and naturally create mental models of their opponents, inferring their goals, and constructing their strategies around behaviors that they may not share themselves.

4 Methods

We next describe the methods employed in this work, detailing the training of a Deep Q Network (DQN) to learn a favorable strategy for the game. We also discuss the difficulty in interpreting the learned strategy and our objective to uncover whether the DQN models its opponent (constituting TT) or simply assumes a specific policy (constituting ST).

4.1 Training the Deep Q Network

We trained a Deep Q Network (DQN) to play the Bracket Stamina game, aiming to learn a strategy that would significantly outperform a random policy. The DQN was trained using a combination of experience replay and target network updating. Over the course of training, the DQN learned a strategy that outperformed an equivalent random policy by more than double (an 8.1% gain), equivalent to a random policy with roughly 8 more points than the other agents.

However, interpreting the learned strategy proved challenging. Our goal is to uncover whether the DQN models its opponent (in line with TT) or if it simply assumes a specific policy, either its own or a random one (in line with ST). To achieve this, we analyze the neural network's representations and decisions to determine if it demonstrates a rich understanding of its opponent's decision-making processes or merely infers a single specific policy. What we found was that it learned a relatively round-agnostic "phase diagram", suggesting it was learning a general strategy for the structure of the game rather than any particular behaviors of its opponents.

5 Experiments

In this section, we run some example games with DQN agents, human agents, and a simple consistent proportion heuristic agent. Description of these agents can be found below:

- **DQN** A Deep Q Network (DQN) which is trained against random policy agents and learns to win with higher probability. To make decisions in a continuous state space, the model

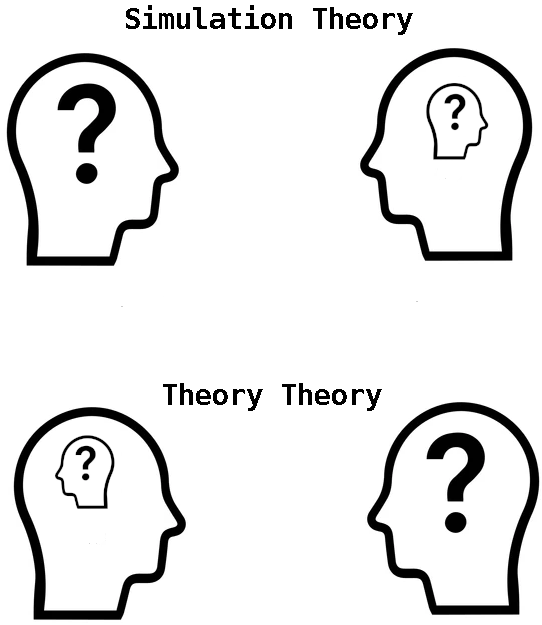

Figure 2: Simulation Theory vs Theory Theory. The crux of our work is a setting where traditional reinforcement learning paradigms (e.g. self-play) fail to generalize to settings with mixed strategies. Simulation theory is the dominant strategy in traditional RL settings, but our work proposes a setting where the latter may be required for strong performance against many distinct agent types.

would instead output a percentage of its remaining points to wager, which would discretize the state space into 100 bins rather than any floating point value. The model was trained in PyTorch [13] on a single Nvidia 1080 Ti.

- **Human Opponents** Human opponents. Participants were selected from a large gaming discord among players who already understood elimination tournaments very well.
- **Consistent Proportion Heuristic** A simple heuristic model loosely based on the Kelly criterion. Select a random value between 4 and 9, and then use that proportion of your remaining wealth every round until the final round, where you use all your remaining wealth.

Diagrams for these tournaments can be found in the appendix. Overall, the consistent proportion heuristic placed below-average compared to the human players, the DQN placed about average compared to human players, and the best human players consistently and significantly outperformed the DQN models.

Two participants in particular consistently placed above the norm compared to the other participants. When interviewed, these players specifically (1) simple "general" heuristics of betting a proportion of wealth, and (2) building mental profiles of their opponents, and specifically adjusting their simple heuristics around defeating them. An analogy can be drawn to Wang et al. [19], where they describe a weak agent which can defeat superhuman go AI through an adversarial policy which does not work on human opponents. As the players became more familiar with how the bot sized its bets, the bots began performing substantially worse than human players.

6 Discussion

In this work, we introduced the Bracket Stamina game as a novel multiplayer generalization of the Kelly Coin Flip game, which requires players to strategically manage their resources while inferring their opponents' decision-making processes. We designed and trained a Deep Q Network (DQN) to play this game, which successfully learned a strategy that outperformed an equivalent random policy by more than double, given the same resources. However, our experiments showed that the DQN

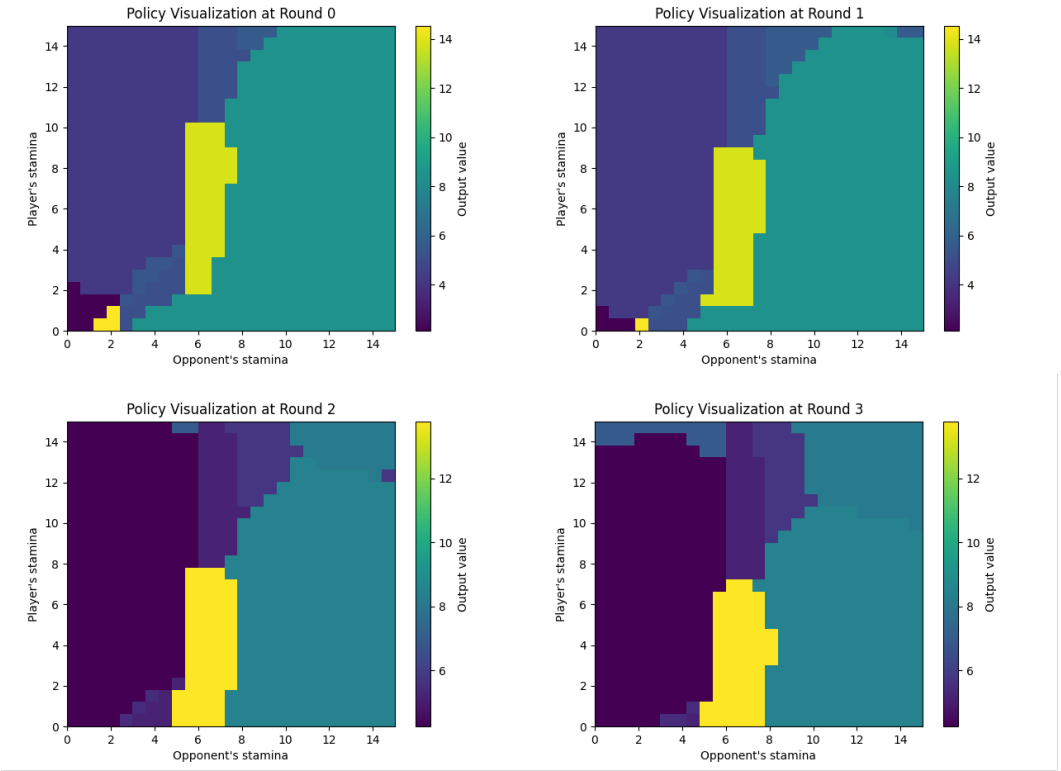

Figure 3: The apparent decision-making process of our learned DQN model. While it reliably defeats random policies at a much higher chance than random, it has some confusing emergent decision-making. For instance, it overbets in the "phase transition" period where it transitions from a low bet to a moderate bet, and it seems to never learn that it should spend it's entire budget in the 4th round.

did not generalize well when confronted with opponents from different decision-making processes, indicating that it did not develop a rich representation of its opponents' strategies.

Our experiments reveal in interesting nuance in the delineation between ST and TT – namely, that decision-making processes which are constructed based on other agents' behaviors are not necessarily directly projecting the agent's own policy into the other player. In this experiment, the DQN clearly assumed all of it's opponents would be using a random policy, and it's own policy differed substantially from that of a random policy. This raises the question: does ST necessarily require projecting your own decision-making process to other agents, or is it more generally projecting your representation of "agent policies" in general which constitutes ST? Is TT a perspective associated with perspective-taking, or higher computational sophistication? If a model learns a rich enough distribution of agent archetypes, and learns to identify them well, does that sufficiently advanced ST begin to resemble TT? The line between these two descriptors of theory of mind are perhaps more nebulous than initially hypothesized.

Much in the same way that Rapoport et al. [15] demonstrated that a varying distribution of agents in a tournament could yield different answers for a theoretically "optimal" strategy, so too does it seem possible that the DQN could have potentially learned a policy which hedged against the existence of other policies in a given tournament (i.e. not all random agents). In this scenario, it could theoretically emerge slightly more robust to the existence of agents of particular types, compared to its current form where it will assume its opponent will make a uniformly random selection. However, important to note is that this new policy conditioned on other agents may fail to generalize to other tournaments much in the same way tit-for-tat failed to generalize in Rapoport et al. [15]. This highlights yet another weakness of the DQN approach – that it's policy is fixed, contingent upon an a priori distribution of

adversarial policies, and that it cannot adjust it's strategy in repeated games the way that humans or learning-enabled agents can.

Our findings highlight the limitations of the current DQN model in adapting to diverse opponent decision-making processes. This observation has broader implications for the understanding of large language models (LLMs), as it suggests that LLMs may also struggle with modeling the intentions and strategies of others when confronted with diverse or unfamiliar decision-making processes. A better understanding of the underlying "psychology" of neural networks could lead to the development of more robust and adaptive artificial agents, capable of effectively operating in complex, socially-driven environments.

There are several possible directions to address the limitations observed in our study. First, incorporating meta-learning techniques into the DQN model may enable it to adapt more quickly to new opponents and decision-making processes. This would allow the network to learn a more flexible representation of its opponents' strategies, potentially leading to better generalization.

Another possible direction is to incorporate elements of ST and TT directly into the architecture of the neural network. This could involve designing specialized modules within the network that explicitly model the intentions and decision-making processes of opponents, enabling the network to make more informed predictions about their actions.

Finally, the use of unsupervised learning and representation learning techniques may help uncover the latent structure in the decision-making processes of opponents. By learning to represent these processes in a more structured and interpretable way, the neural network could better adapt its strategy to diverse opponent types.

By exploring these avenues, we hope to gain a deeper understanding of the emergent decision-making processes in neural networks, particularly in the context of large language models. This could not only lead to more adaptive and effective artificial agents but also provide insights into the otherwise non-interpretable representations of others in LLMs.

Acknowledgments and Disclosure of Funding

This paper was completed with no external sources of funding.

References

[1] Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24(1):3–25, 1980.

[2] Gwern Branwen, Arthur Breitman, nshepperd, FeepingCreature, and Gurkenglas. The kelly coin-flipping game: Exact solutions. 2017.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[4] Siang Yew Chong, Jan Humble, Graham Kendall, Jiawei Li, Xin Yao, et al. The iterated prisoner's dilemma: 20 years on. *Advances in Natural Competition*, 4:1–22, 2007.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

[7] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998.

[8] Alison Gopnik and Henry M Wellman. The theory theory. In *An earlier version of this chapter was presented at the Society for Research in Child Development Meeting, 1991.* Cambridge University Press, 1994.

[9] Victor Haghani and Richard Dewey. Rational decision-making under uncertainty: Observed betting patterns on a biased coin. *arXiv preprint arXiv:1701.01427*, 2017.

[10] John L Kelly. A new interpretation of information rate. *the bell system technical journal*, 35(4):917–926, 1956.

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[12] John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[15] Amnon Rapoport, Darryl A Seale, and Andrew M Colman. Is tit-for-tat the answer? on the conclusions drawn from axelrod's tournaments. *PloS one*, 10(7):e0134128, 2015.

[16] Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.

[17] Tuomas W Sandholm and Robert H Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1-2):147–166, 1996.

[18] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[19] Tony Tong Wang, Adam Gleave, Nora Belrose, Tom Tseng, Joseph Miller, Michael D Dennis, Kellin Pelrine34, Yawen Duan, Viktor Pogrebniak, Sergey Levine, et al. Adversarial policies beat superhuman go ais.

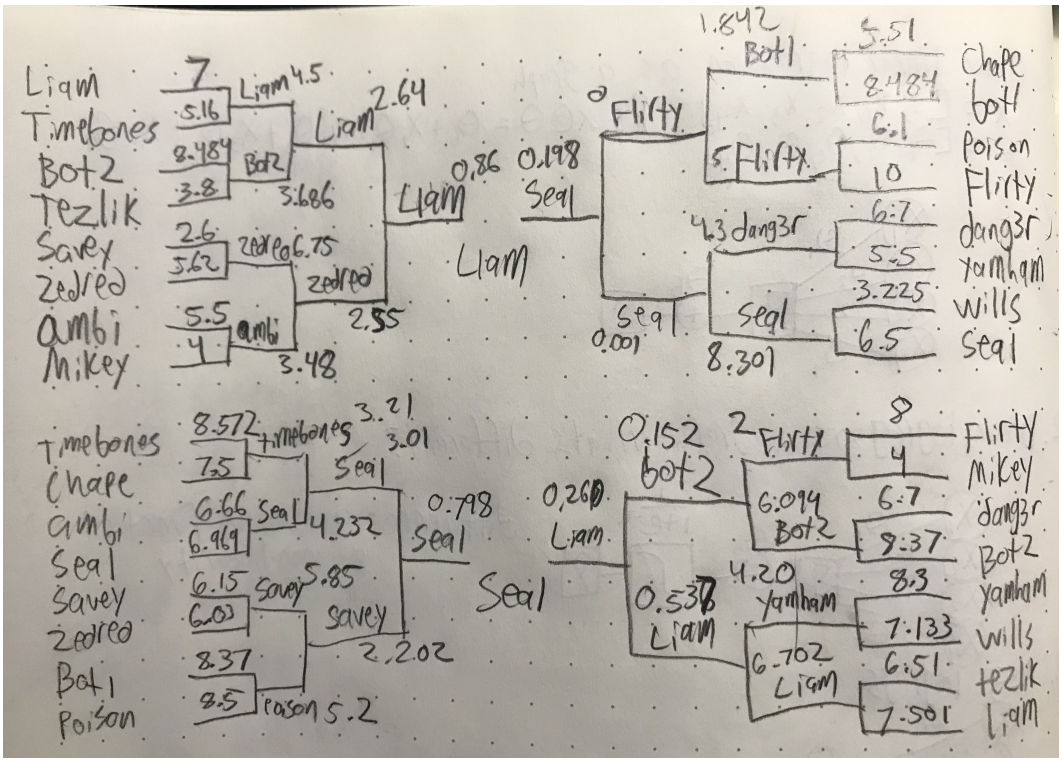

Figure 4: results of human-machine mixed games. This figure will be replaced with a higher resolution example further into the future.

A Further Details

On this page some example games with both human and machine agents can be found.