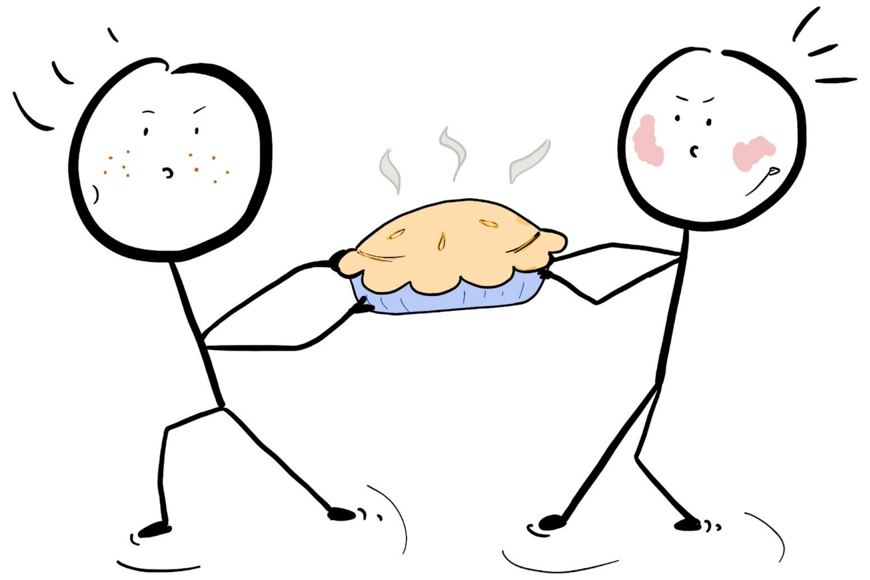


Daphne Cornelisse[1], Thomas Rood[1], Mateusz Malinowski[3], Yoram Bachrach[3], Tal Kachman[1, 2]

1. Department of Artificial Intelligence, Radboud University, Netherlands
2. Donders Institute for Brain, Cognition and Behavior, Radboud University, Netherlands
3. DeepMind, UK

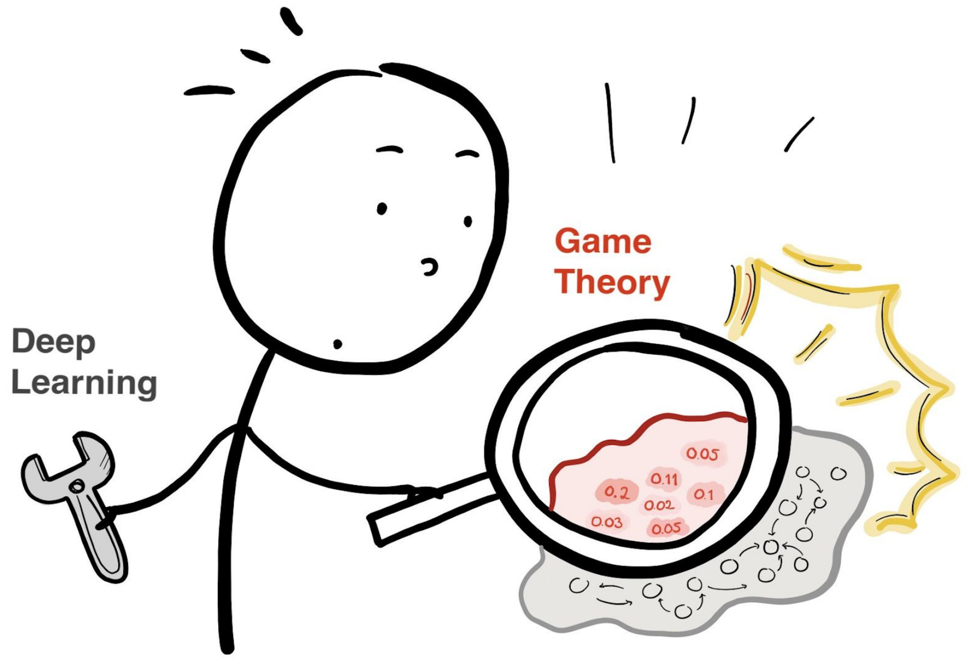
Radboud University

NEURAL INFORMATION PROCESSING SYSTEMS

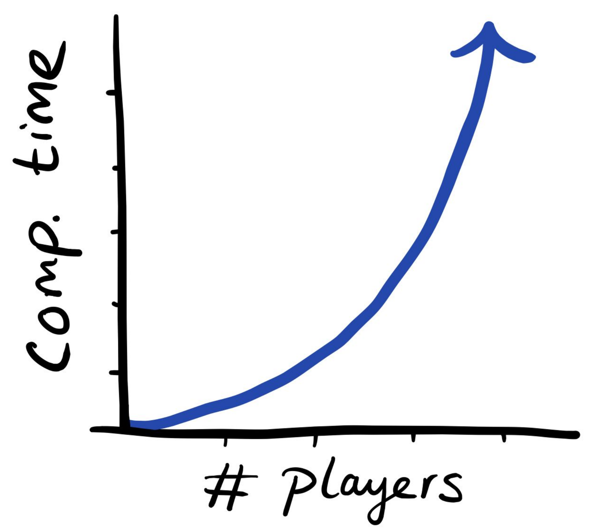
DeepMind

Neural payoff machines: Predicting *fair* and *stable* payoff allocations among team members

Motivation. How can we understand the *contribution* of single agents in large multi-agent systems?



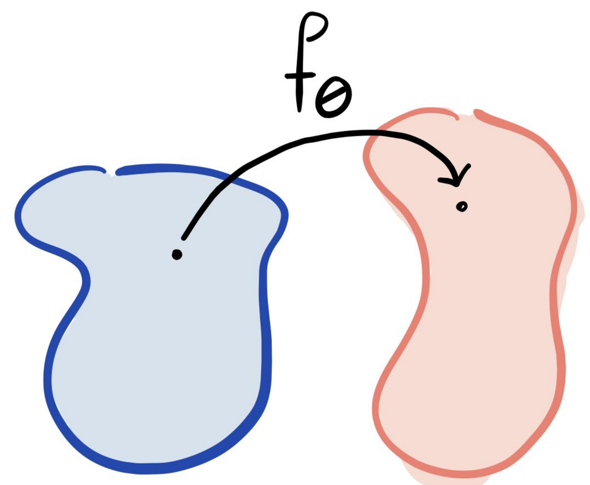
Problem. Game theory provides useful formalisms, such as notions such as **fairness** and **stability**, but these are **NP-hard**.



Research Questions.

1. Can **stability** and **fairness** metrics be learned in multi-player games?

2. Do models trained for **small** games generalize to **larger** games?



Background

Game class: Weighted voting games

Set of players $N = \{1, 2, \ldots, n\}$

Set of weights $\mathbf{w} = (w_1, w_2, \ldots, w_n)$

Quota (threshold) q

Characteristic function $v : 2^N \to \{0, 1\}$

- Maps each coalition $C \subseteq N$ to a value $v(C)$

Solution concepts

Fairness

Shapley value

$$\phi_i(G) = \frac{1}{N!} \sum_{\pi \in \Pi} [v(C_\pi(i) \cup \{i\}) - v(C_\pi(i))]$$

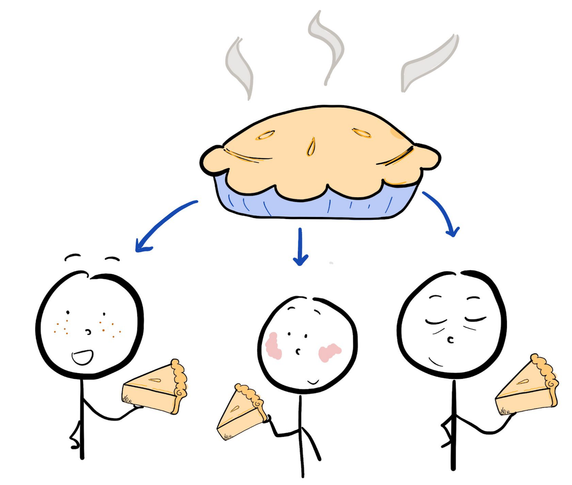
Banzhaf index

$$\beta_i(G) = \frac{1}{2^{N-1}} \sum_{C \subseteq N \setminus \{i\}} [v(C \cup \{i\}) - v(C)]$$

Stability

The least core

$$\begin{aligned} \min_{p_1, p_2, \ldots, p_n, \varepsilon} \quad & \varepsilon \\ \text{s.t.} \quad & \sum_{i \in C} p_i \geq 1 - \varepsilon \quad \forall C \in C^{\min} \\ & \sum_{i \in N} p_i = 1 \\ & p_i \geq 0 \quad \forall i \in N \end{aligned}$$



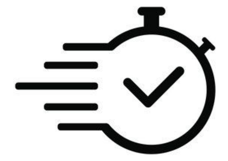
$$C^{\min} \equiv \left\{ C \subseteq N \;\middle|\; C \neq \varnothing, \sum_{i \in S} w_i \geq q, \text{ and, for all } j \in C, \sum_{i \in C \setminus \{j\}} w_i < q \right\}$$

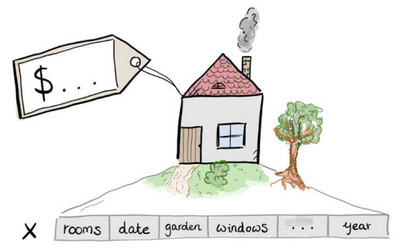
Results & Conclusions

1. Fairness & **stability** metrics can be learned. Stability is harder.

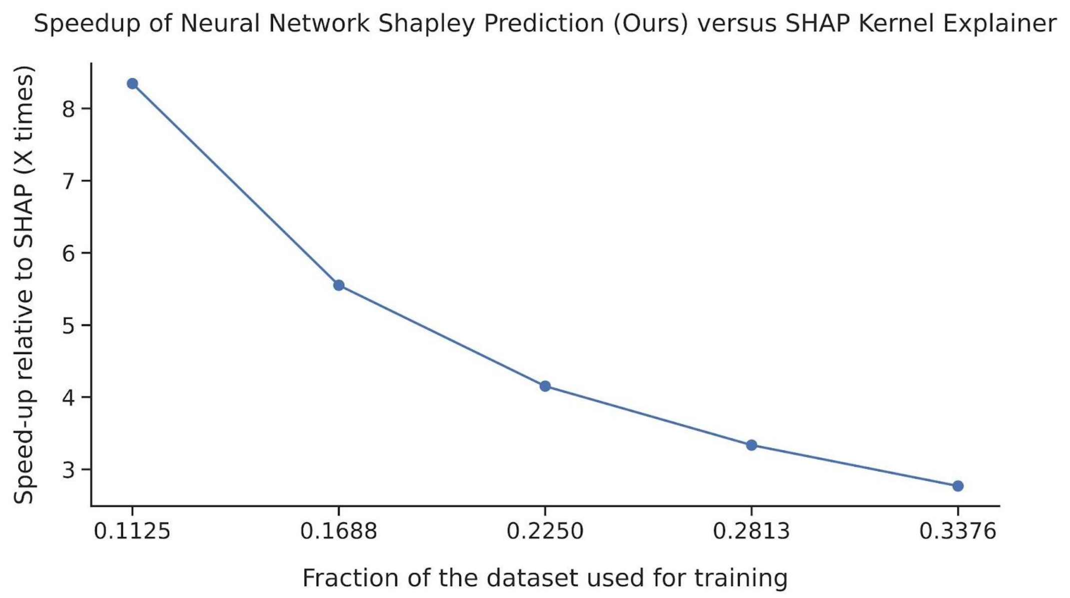
Table 1: Comparison of predictive performance across test datasets and solution concepts.

	Least core payoffs		**Least core excess**		**Shapley values**		**Banzhaf indices**	
Dataset	Mean MAE		MAE		Mean MAE		Mean MAE	
	Fixed	Variable	Fixed	Variable	Fixed	Variable	Fixed	Variable
In-sample	0.030	0.043	0.015	0.034	0.019	0.022	0.018	0.028
Out-of-sample	0.030	0.044	0.015	0.027	0.018	0.036	0.018	0.056
Slightly out-of-distribution	0.029	0.028	0.015	0.050	0.018	0.019	0.017	0.018
Moderately out-of-distribution	0.030	0.035	0.014	0.029	0.018	0.026	0.018	0.032
Significantly out-of-distribution	0.031	0.045	0.014	0.036	0.018	0.029	0.018	0.039

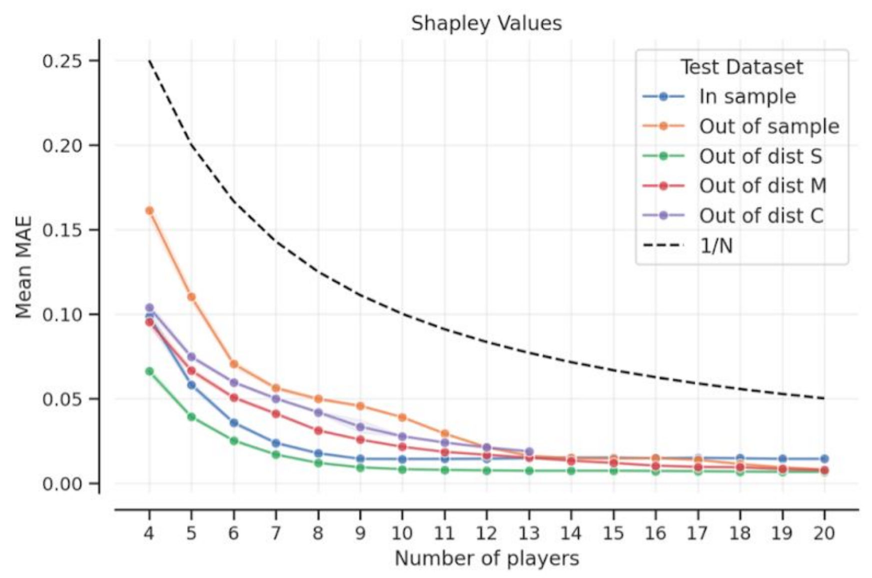


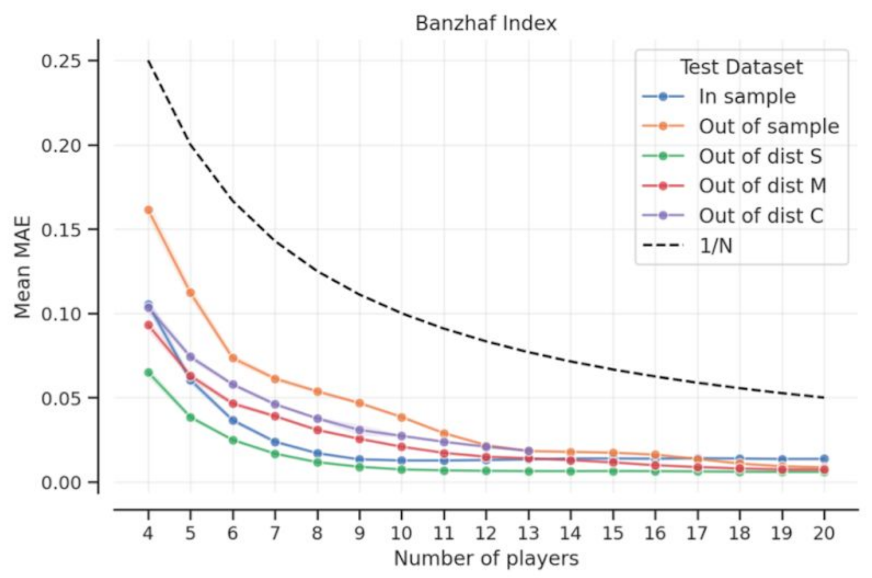


Application: > 8x Speedup explainable AI analysis



2. Models **generalize** to **larger** games to some extent (small decrease in performance as $n > 10$).





References

E. Elkind, L. A. Goldberg, P. W. Goldberg, and M. Wooldridge. *Computational complexity of weighted threshold games.* In AAAI, pages 718–723, 2007

X. Deng and C. H. Papadimitriou. *On the complexity of cooperative solution concepts.* Mathematics of operations research, 19(2):257–266, 1994.