SUBJECT: AI to Support Humans in Imperfect Information Games

SUPERVISOR: Ph. D., DSc, Eng. Agnieszka Pregowska, Institute of Fundamental Technological Research, Pawińskiego 5B, 02–106 Warsaw, PL, e-mail: aprego@ippt.pan.pl

Name of the institute in which the topic will be realized: **Institute of Fundamental Technological Research** Scientific discipline: **Information and communication technology**

Description

Artificial intelligence has achieved remarkable performance in complex strategic games, including environments characterized by large state spaces and imperfect information [1-3]. However, real-world strategic decision-making differs substantially from conventional game solving. Opponents may be boundedly rational, adaptive, predictable or deliberately deceptive, while their objectives, information and future actions are only partially observable. Consequently, an AI system intended to support human decision-makers should not only identify strong strategies but also model the behaviour of other agents and remain robust when these models are uncertain or incorrect.

The aim of this PhD project is to develop and evaluate AI methods for strategic decision support in multi-agent environments with imperfect information. Particular emphasis will be placed on opponent modelling and safe opponent exploitation, using predictable patterns in an opponent's behaviour while limiting vulnerability to modelling errors, behavioural changes and deliberately misleading actions [4–6].

The research will combine reinforcement learning, multi-agent reinforcement learning and computational game theory. Equilibrium-oriented methods, counterfactual regret minimization, search-based algorithms and reinforcement-learning policies will provide reference approaches. The main research challenge will be to develop adaptive methods that incorporate information about a specific opponent while maintaining robustness when the assumed opponent model becomes inaccurate.

The project will address the following research problems:

- modelling opponent behaviour under partial observability;
- integration of opponent models with learning and game-theoretic decision algorithms;
- adaptive exploitation of predictable or boundedly rational opponents;
- robustness to incorrect opponent models, behavioural changes and deceptive strategies;
- quantitative analysis of the trade-off between opponent-specific adaptation and worst-case performance;
- analysis of AI-agent behaviour relevant to human strategic decision support.

The proposed methods will be evaluated in controlled imperfect-information game environments involving sequential decisions, hidden information and competing agents. Performance measures will include expected return, win rate, prediction accuracy of opponent actions, exploitability and robustness to changes in opponent behaviour.

An additional research dimension will concern the relationship between computational strategies and human decision-making. Recent studies indicate that strategic AI should account not only for formal game performance but also for human compatibility, calibrated trust and the limitations of automated decision support [7,8]. Concepts such as bounded rationality, trust and deception may therefore be incorporated into opponent models and experimental scenarios.

The expected outcome of the project is a new methodological framework for robust and opponent-aware strategic AI, capable of exploiting useful information about other agents without excessive loss of robustness. The ultimate objective is to develop AI methods that complement human strategic reasoning rather than replace it with opaque automated decisions.

Selected references

[1] Brown, N., Sandholm, T., "Superhuman AI for multiplayer poker," *Science*, 365, 885–890, 2019.

[2] Perolat, J. et al., "Mastering the game of Stratego with model-free multiagent reinforcement learning," *Science*, 378, 990–996, 2022.

[3] Rudnicka Z., Szczepanski J., Pregowska A., Accuracy-efficiency trade-offs in spiking neural networks: A Lempel-Ziv complexity perspective on learning rules, Applied Soft Computing 203, 116012, 1-15, 2026.

[4] Milec, D. et al., "Complexity and Algorithms for Exploiting Quantal Opponents in Large Two-Player Games," *Proceedings of AAAI*, 2020.

[5] Liu, M. et al., "Safe Opponent-Exploitation Subgame Refinement," *Advances in Neural Information Processing Systems*, 35, 2022.

[6] Ge, Z. et al., "Safe and Robust Subgame Exploitation in Imperfect Information Games," *Proceedings of ICML*, 2024.

[7] Bakhtin, A. et al., "Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning," 2022.

[8] Tomsett, R. J. et al., "Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI," *Patterns*, 1, 100049, 2020.