9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Designing Human-Compatible AI: Methodologies and Frameworks in Chess

Keywords: Human-AI compatibility, Agent Systems, Decision-making, Chess, Deep RL

Artificial intelligence has long achieved superhuman performance in many decision-making tasks, and state-of-the-art AI dominates professionals in games like chess and GO. At the same time, AI decision-systems are becoming increasingly prevalent in real use-cases where they frequently interact with humans. For AI agents that interact with humans in complex environments, like driving, to become widespread, achieving superhuman performance is not sufficient: agents need to account for human actions from human counterparts. Our work uses chess (specially designed variations thereof to include a cooperative aspect with humans) as a model system to study artificial intelligence agents that dynamically interact with humans. Traditional chess engines designed to output near-optimal moves prove to be less adequate as advisors for humans of different skills, with their move suggestions often complicated and difficult to build on. Our contribution is therefore the proposition of three methodologies to create human-compatible AI in a complex decision-making system as chess, as well as the evaluation of these methodologies on two chess-based frameworks designed to simulate and foster human-AI collaboration.

We select chess as our model system due to the ready availability of both superhuman AI agents and AI agents designed to emulate human players, the complexity of the task, and the availability of large amounts of human data. The superhuman engines we will be using throughout this paper are the AlphaZero based Leela engines [2], and the human emulators are the Maia engines [1]. The Maia engines, trained on tens of millions of human games of different skills, will act as proxies of humans, a critical innovation necessary to play the tens of thousands of games needed for the creation and testing of our new agents. While they are not perfect embodiments of human players, they are the state-of-the-art engines at predicting human play, and they capture human-like patterns of idiosyncratically sub-optimal play.

We propose the following chess frameworks:

Coin Flip framework: This setup consists of two teams, each team in control of a color on the chessboard. A team consists of two agents, the junior teammate, which will generally be Maia, and the senior teammate which will generally be a stronger engine. At every turn, the team whose turn it is flips a fair coin, with the outcome determining whether the senior or junior takes to the board to play. The setting introduces a cooperative aspect to chess, meaning a senior will need to ideally be prepared for the possibility that a weaker junior might be making the next move. At the same time, the senior will need to play at a high level of chess, as the opponent senior will be playing at a high level. This framework is designed to emulate situations where allies and opponents can be both AI and human, and the AI is required to both perform at a high level, and account for human involvement. See figure 1

Hands and Brains framework: this setup consists of two teams, each team in control of a color on the chessboard. A team consists of two agents, the brain agent (always the stronger agent in our case), which selects the piece type to be moved (ex: knight), and then the hand (Maia agent in our case), which then selects the exact piece and move to make(ex: knight from g1 to f3). In contrast to the previous framework, this framework emulates cases where the AI narrows the action-space for humans, who make the final decision.

We design 3 novel methodologies that leverage Maia to create human-compatible engines:

Tree agent: By using the Maia models as the basis of a tree search, we design a new agent that attempts to anticipate Maia's choices. This agent is notable in that it only requires a relatively accurate model for human play, without the use of a superhuman agent.

Expector agents: These agents maximize the expected value of the game, taking into account Maia intervention, over a short time horizon. They internally need a strong evaluating agent (we will use Leela), and a model for the allied junior/hand (as well as for the opponent junior in the coin flip framework.)

Trained agents: These agents are modified versions of Leela, which is taken and made to play versus itself in one of the frameworks as a senior/brain, with Maia being the junior/hand. The games are then used to modify the weights of Leela to take account for the involvement of Maia.

The main evaluation consists of deploying the above agents in the frameworks as seniors/brains, versus Leela as senior/brain on the other team, with Maia as junior/hand for both teams. We report the expected score over 100 games. The score is defined as the difference between the number of wins for our agents and number of losses for our agents. The results are shown in table 1. We see that all 3 agent classes achieve a positive expected score on both frameworks. The last column shows that none of the agents are able to beat Leela in a conventional chess game, with no Maia involvement. The table thus encapsulates the principal result of our work: the designed agents, inferior to Leela conventionally, outperform it in the presence of Maia agents (both as a hand, and as a junior). This demonstrates the feasibility, and necessity of sacrificing optimality to account for human compatibility. Our full results explore and demonstrate the generalizability of these methods to different juniors, and explore the interactions between the seniors/brains and juniors/hands in a series of ablation tests. Future directions of work include validating our models in settings with actual humans, as well as the generalization of our techniques beyond the game of chess and to more involved and realistic settings.

Agent Class/ Type of Game	Coin Flip (± 1)	Hands and Brains (± 1)	Conventional Match (± 10)
Tree	13.5	7.9	-99
Expector	32.8	18.1	-72
Trained	9.6	22.35	-75

Table 1: Expected Score over 100 games of the different agents versus Leela in different frameworks. Maia is used as the junior and brain where applicable, with standard deviation reported

References

[1] Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman ai with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1677–1687, 2020.

[2] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

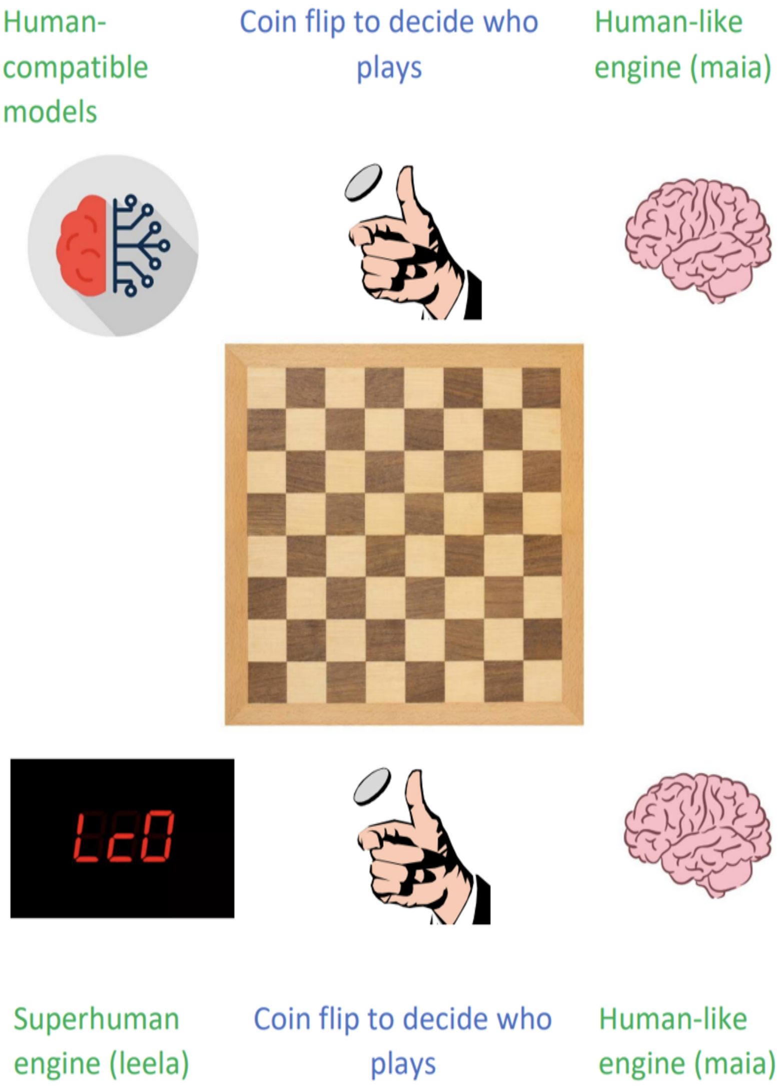
Human-compatible models
Coin flip to decide who plays
Human-like engine (maia)
LcO
Superhuman engine (leela)
Coin flip to decide who plays
Human-like engine (maia)

Figure 1: Illustration of the teams in the Coin Flip Framework