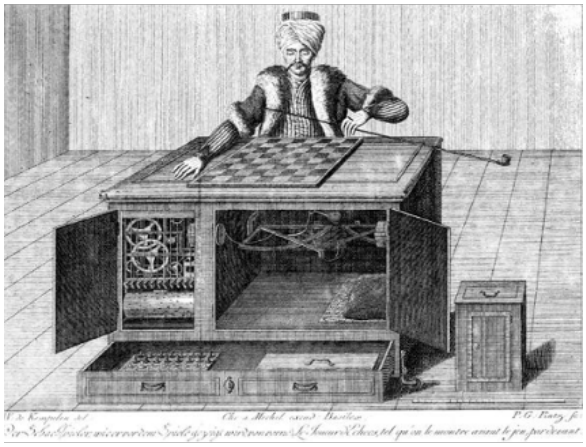
AI at the Chessboard:

Predicting, Understanding, and Aligning with Human Thought

Paper written by:
Anthony **Carbonelli**
Ksenija **Kankaraš**

HEPHAESTUS APPLIED ARTIFICIAL INTELLIGENCE ASSOCIATION

When a man sits at the chessboard, he carries his excitement and stress, leftover thoughts of the day, and all his past experiences along. No such burdens are programmed in the AI system playing against him; it possesses the advantage of focused analysis, without human distractions holding it back.
The first attempt at automating chess was in 1770, when a Hungarian inventor, Kempelen, presented the Mechanical Turk. Although it was merely a box with a chess master hidden inside controlling the moves, it fooled audiences around Europe for nearly nine decades. People's fascination planted the seed of ideas. What if a machine could master the art of chess?



Numerous attempts at this followed throughout history, such as El Ajedrecista (The Chessplayer), and Turochamp by Alan Turing. The computer engines of the 1960s and 1970s were unable to compete successfully with chess masters. A turning point occurred in 1995 when a new chess engine prototype, Deep Blue, was released by International Business Machines Corporation (IBM).
After facing Deep Blue in 1997, the World Chess Champion at the time, Garry Kasparov, claimed that he noticed several moments in which the AI exhibited human-like creativity. With a FIDE (International Chess Federation) score of 2800 and a streak of 12 world chess titles in a row, Kasparov was considered the greatest chess player in history going into his match with Deep Blue. As the match progressed, Kasparov claimed to have "lost his fighting spirit" and resigned against the computer. This raises the question: Do AI systems understand human concepts, or do they utilize pattern recognition systems that happen to produce these creative outputs?
The impracticality of trying to "solve chess" by exhaustive search was explained by Claude Shannon in 1950, in his book "Programming a Computer for Playing Chess", by calculating a lower bound of the game-tree complexity of chess, resulting in about 10120 possible games. Nevertheless, its complexity is the reason it was initially perceived as an ideal testbed for the development of computational reasoning, and why it is well-suited for the pursuit of clarity on human-AI alignment today.
With the rise of neural network analyses, engines have become more advanced than researchers could have imagined. While "superhuman" AI chess engines like Stockfish and AlphaZero are realistically undefeatable by humans, we cannot comprehend their methods and learn from them. This partially led to the development of Maia – a collection of models, mimicking human behavior in chess at various skill levels, by using a wealth of online gameplay data to predict actual human moves.

On a similar note, recent research on Artificial Intelligence in Chess demonstrates the abilities of modern-day machines to use elegant algorithms and models to predict an individual's next move based on their skill level, to learn advanced patterns that simulate the understanding of human concepts, and finally, to "perfect" the imperfect human behavior in the match.

One of the most intriguing approaches comes from the study using skill-group-specific n-gram models. Instead of trying to capture universal chess logic, the researchers trained models separately on games from different rating ranges. This instantly revealed something both obvious and profound: a beginner's mental landscape is not simply a weaker version of a master's, but a fundamentally different one. Predictive accuracy jumped once models were tailored to the distinctive habits of specific skill groups. What emerges is a portrait of chess cognition as a sequence-learning process in which humans rely on familiar patterns and local motifs, often without consciously articulating them.

Another strand of research turns inward, examining the internal activations of neural chess networks. The study on look-ahead behaviour demonstrated that these networks seem to build implicit multi-step predictions even without an explicit search module. As inputs proceed deeper into the network, the representations begin to resemble future board states. The model is, in a sense, learning to visualize. This bridges an unexpected gap between human and machine reasoning: both appear to solve chess by mentally simulating the near future, though they do so through entirely different mechanisms.

The conceptual-alignment study tackles a broader and more philosophical problem. Human players describe positions using a vocabulary of ideas, weak squares, initiative, king safety and compensation, which have no direct analogue in the numerical tensors of a neural network. The researchers attempted to map the geometry of these internal representations onto human conceptual categories. Some concepts aligned cleanly; others were more diffuse or fragmented across the model's internal space. This ambiguity speaks to a deeper challenge: even when an AI's behaviour resembles human thought, its internal logic may be organized along alien dimensions.

Maia-2 closes the circle by operationalizing alignment rather than merely studying it. Built specifically to predict human moves, it does not seek perfect play but a faithful imitation of how humans actually behave. It learns not only the optimal line, but the common mistakes, the hesitations, the moments when a player chooses a comfortable idea over a materially superior one. By doing so, Maia-2 becomes a lens on human cognition and a tool that can teach, analyze, and even empathize with human decision processes.

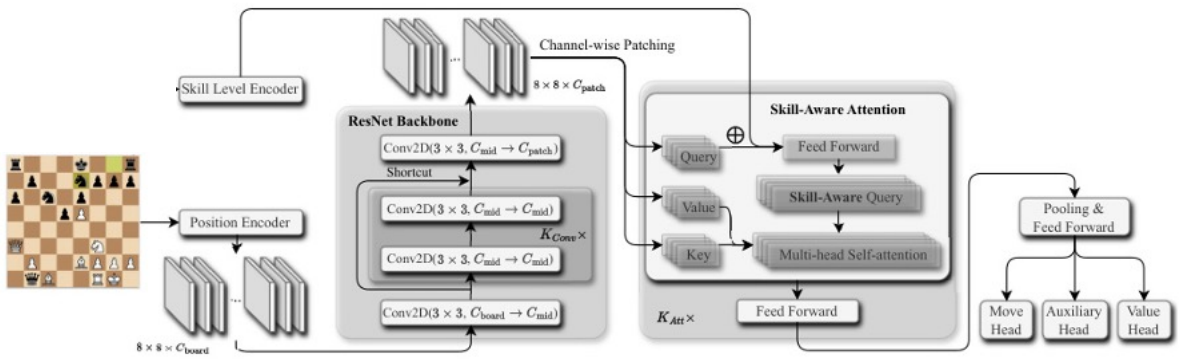


Figure 1: *Skill-aware neural architecture for human move prediction.*

Despite their different aims, these studies share several core techniques. They treat chess games as structured sequences, allowing statistical models to learn the equivalent of a "chess grammar." They probe neural representations with techniques reminiscent of cognitive neuroscience, dissecting embedding spaces and activation patterns to infer how models organize information internally. They condition their training on human skill levels so that the model learns the specific heuristics characteristic of each group. They reveal that even without explicit search algorithms, deep networks naturally evolve internal forms of forward reasoning. And they evaluate success not only by comparing moves to optimal play, but by asking how closely models match the choices, thought patterns, and conceptual structures of actual human players. All of this marks a departure from traditional engine design, pushing AI research toward a more interpretive and psychological direction.

Yet the path toward fully understanding human chess reasoning remains long. Predictive models can absorb patterns but cannot yet capture the full texture of human cognition. A human's decisions are shaped by emotion, fatigue, prior experience, and narrative thinking, factors that leave no direct trace in a dataset of moves. Conceptual alignment, even when statistically strong, does not guarantee that a model truly "understands" a concept in the human sense. It may merely correlate with it.
There is also a subtle divergence between superhuman logic and human heuristics that no amount of alignment can fully erase. An AI may describe its evaluation in familiar language, but internally it might rely on structures that have no intuitive counterpart. Move prediction itself remains limited: humans blunder, improvise, and act inconsistently, making their choices inherently noisy.

Finally, there are ethical questions that hover around such research. As these models become increasingly adept at predicting human behaviour, even in a structured domain like chess, it forces us to confront similar possibilities in domains where prediction touches privacy, autonomy, or competitive fairness. An engine that knows a player's tendencies better than the player does might be a teaching tool, or a manipulative one.

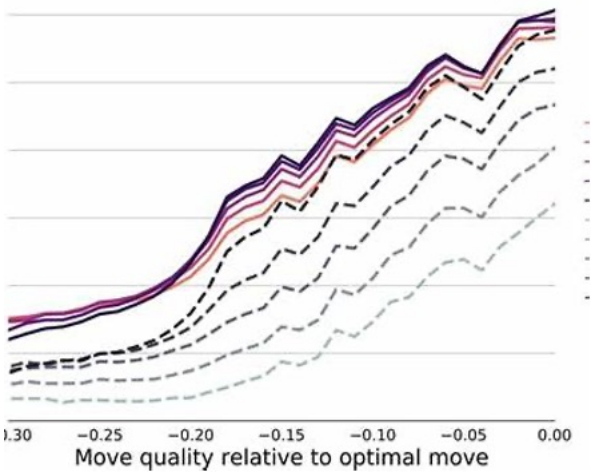


Figure 2: Relationship between human move quality and deviation from optimal engine play.

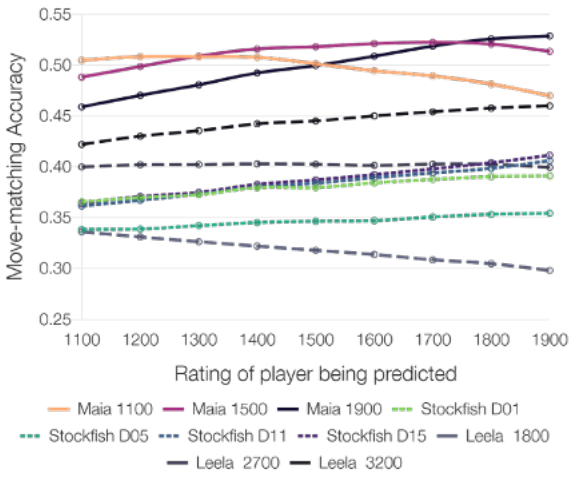


Figure 3: Move-matching accuracy of Maia models versus traditional engines across player rating levels.

What emerges from these studies is a portrait of AI research migrating away from the quest for mechanical perfection and toward a deeper engagement with human thought. Chess, in this context, becomes less a battleground and more a dialogue between two kinds of minds, one biological, one computational, each learning to understand the other. Models built to mirror human play reveal patterns that humans rarely articulate. Networks that implicitly look ahead hint at forms of reasoning that converge with human intuition. Conceptual-alignment studies highlight the delicate boundary where machine representations begin to resemble our own conceptual frameworks. Maia-2 demonstrates that a chess engine can align itself so closely with human behaviour that it becomes a partner rather than a rival.

The broader significance lies in the direction this research points toward. Future AI systems in other domains will need to reason with humans, not merely for them. They will need to communicate in ways that respect human intuition and conceptual structure. They will need to predict human behaviour not to exploit it, but to support it.

Chess, bounded by 64 squares and fixed rules, offers a remarkable proving ground for these aspirations. Through it, researchers are quietly building the foundations of more interpretable, more collaborative, and ultimately more human-aligned artificial intelligence.

SOURCES:

- Predicting Human Chess Moves: An AI Assisted Analysis of Chess Games Using Skill-group Specific n-gram Language Models – arXiv https://arxiv.org/pdf/2512.01880
- Understanding the learned look-ahead behavior of chess neural networks – arXiv https://arxiv.org/pdf/2505.21552
- Exploring Human-AI Conceptual Alignment through the Prism of Chess – arXiv https://arxiv.org/pdf/2510.26025
- Maia-2: A Unified Model for Human-AI Alignment in Chess – arXiv https://arxiv.org/pdf/2409.20553