Aligning Superhuman AI with Human Behavior: Chess as a Model System

Reid McIlroy-Young[1] Siddhartha Sen[2]
Jon Kleinberg[3] Ashton Anderson[1]

[1]Department of Computer Science, University of Toronto

[2]Microsoft Research

[3]Department of Computer Science, Cornell University

July 2020

Motivation

Superhuman AI systems are increasingly prevalent

Some domains humans are moving away from

- facial recognition, path finding, identifying photos of dogs

But other areas will continue see human participation

- poker, chess, some business decisions

Learning from Superhuman AI

In the domains where humans have been superseded but continue to participate, this raises the possibility that we could learn from them

Learning from Superhuman AI

In the domains where humans have been superseded but continue to participate, this raises the possibility that we could learn from them, or collaborate with them.

Superhuman Reasoning

Superhuman AI systems can be difficult for humans to understand, making it difficult for us to interact with or learn from them

Research Question

How can we bridge the gap between the AI's behavior and ours?

Requirements

1. Superhuman AI
2. Observed Humans
3. Diverse Humans

Requirements

1. Superhuman AI
2. Observed Humans
3. Diverse Humans
4. Parametrized Humans

Question

Predict the next move a human, at a specific skill level, will make during a chess game

Lichess

Lichess.org is a popular, free, open-source chess platform with over 1 billion games in its database

ELO Rating

ELO rating is a measurement of skill in a game, the larger the ELO rating the higher the skill

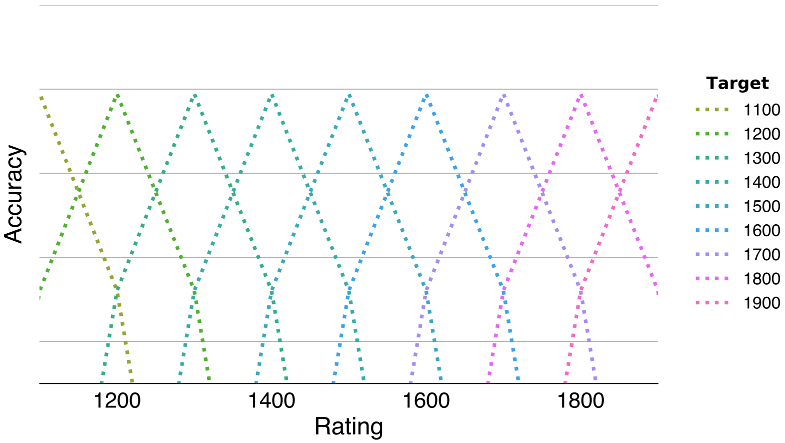
Data

For testing we just use games from December 2019

1. Create bins for each range of 100 rating points
2. Divide games into the bins by the ELO of both players
3. Select 10,000 games from each bin, between 1000 and 2500

For each game we can then look at the mean move prediction accuracy of a model

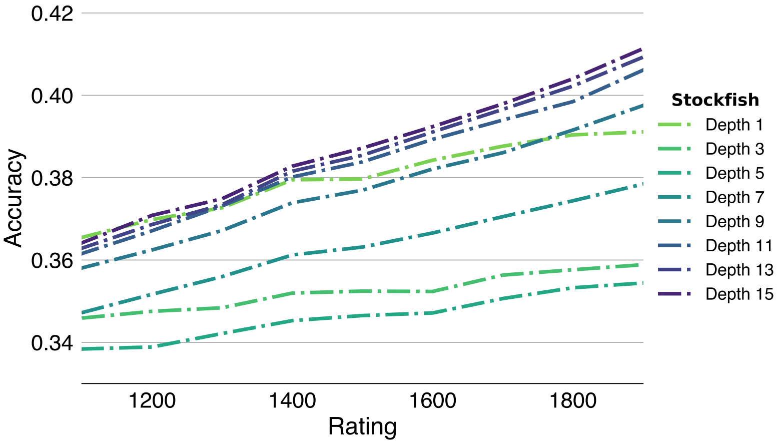
What We Want


Accuracy
Rating
1200
1400
1600
1800
Target
1100
1200
1300
1400
1500
1600
1700
1800
1900

Stockfish Overview

Design	Traditional Chess engine
Type	Tree Search
Humanity	Hand coded heuristics
Tuning	Depth of tree to search

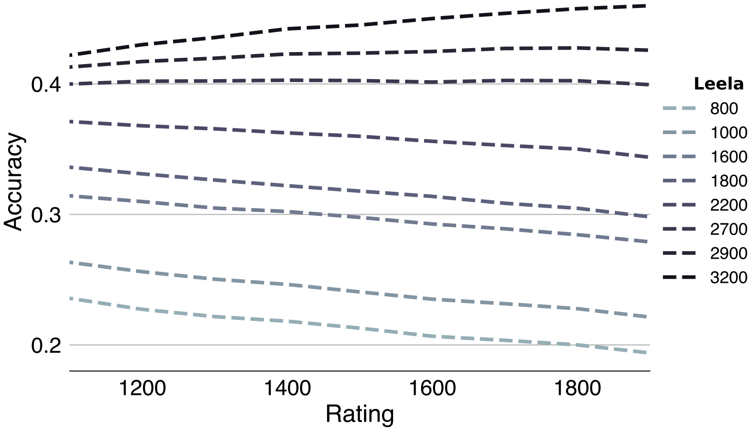
Stockfish


0.42
0.40
0.38
0.36
0.34
Accuracy
1200
1400
1600
1800
Rating
Stockfish
Depth 1
Depth 3
Depth 5
Depth 7
Depth 9
Depth 11
Depth 13
Depth 15

Leela Intro

Design	Implementation of AlphaZero
Type	Reinforcement Learning
Humanity	Only rules of chess
Tuning	Length of training

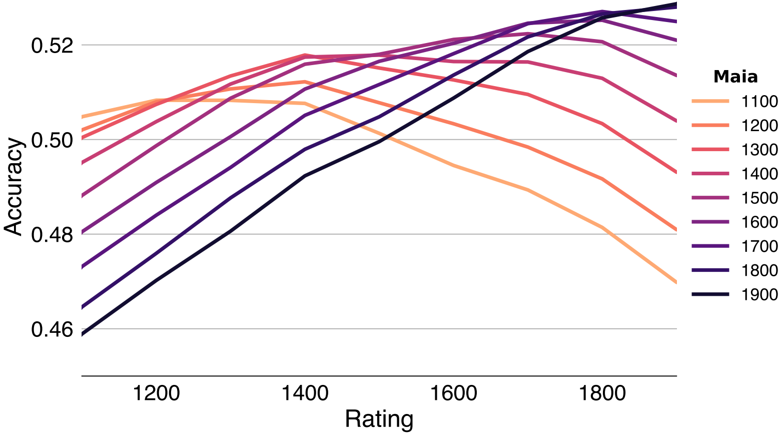
Leela


0.4
0.3
0.2
Accuracy
1200
1400
1600
1800
Rating
Leela
800
1000
1600
1800
2200
2700
2900
3200

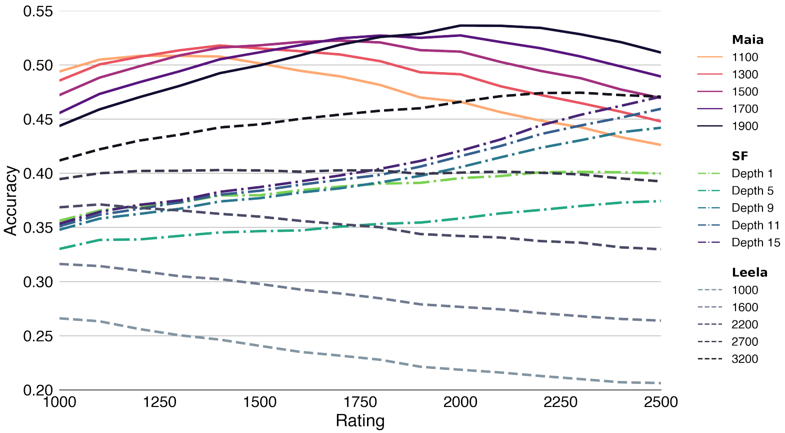
Maia Intro

Design	AlphaZero based deep neural net
Type	Classification
Humanity	Trained on 12 million human games each
Tuning	Trained human skill

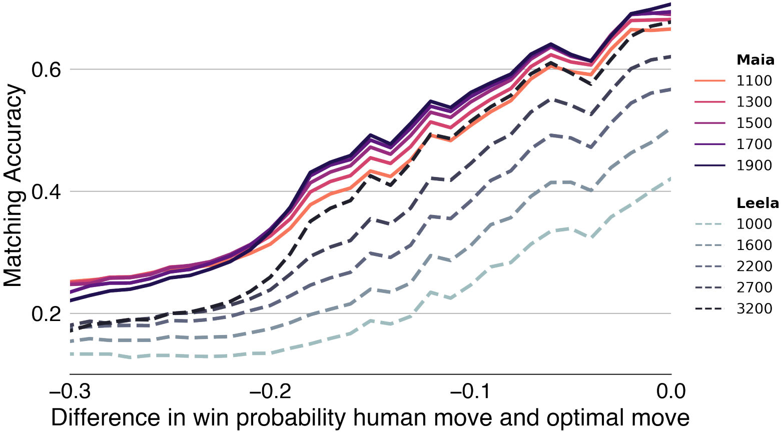
Maia


0.52
0.50
0.48
0.46
Accuracy
1200
1400
1600
1800
Rating
Maia
1100
1200
1300
1400
1500
1600
1700
1800
1900

All


0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Accuracy
1000
1250
1500
1750
2000
2250
2500
Rating
Maia
1100
1300
1500
1700
1900
SF
Depth 1
Depth 5
Depth 9
Depth 11
Depth 15
Leela
1000
1600
2200
2700
3200

Human Errors



Discussion

Mistakes Categorizing

Understanding Human Skill

Learning Aid

Further Information

Paper	KDD 2020
arXiv	`arxiv.org/abs/2006.01855`
Github	`github.com/CSSLab/maia-chess`
Lichess	`maia1`, `maia5`, `maia9`