Chess as a Testing Grounds for the Oracle Approach to AI Safety

James D. Miller, Roman Yampolskiy, Olle Häggström, and Stuart Armstrong

AISafety 2021

To reduce the danger of powerful super-intelligent AIs, we might make the first such AIs oracles that can only send and receive messages. We propose a possible practical means of using machine learning to create two classes of narrow AI oracles that would provide chess advice: those aligned with the player's interest, and those that want the player to lose and give deceptively bad advice. The player would be uncertain which type of oracle it was interacting with. As the oracles would be vastly more intelligent than the player in the domain of chess, experience with these oracles might help us prepare for future artificial general intelligence oracles

Use machine learning techniques to create two classes of chess oracles.

• Friendly oracles would seek to help the chess player win.

• Anti-aligned oracles would give advice that appeared reasonable but was designed to mislead the chess player into making bad moves.

• The player would not know which type of oracle they were receiving advice from.

• Hopefully, playing with such oracles would give us useful and generalizable hints for handling future general intelligence oracles.

• Learning that we cannot reasonably make use of the oracles to improve our chess play would lower the likelihood that the oracle approach to AGI safety is useful.