

chessXAI

<https://forms.gle/hEjPCtDZaka7sSsz8>

Rok temu: AlphaFold święty Graal biologii

THE NOBEL PRIZE IN CHEMISTRY 2024

David Baker "for computational protein design"

Demis Hassabis "for protein structure prediction"

John M. Jumper "for protein structure prediction"

THE ROYAL SWEDISH ACADEMY OF SCIENCES

AI przyjazne i użyteczne dla ludzi

Zakład Humanocentrycznej Sztucznej Inteligencji Uniwersytet Jagielloński

GEIST Research Group

Maciej Szeląg

Maciej Mozolewski

Wyjaśnialność Human-in-the-loop

Zakład HCAI UJ oraz grupa GEIST kierowana przez prof. Cezarego J. Pasieki

Co chcemy zrobić?

Jak objaśnić wnioskowanie?

Czarne pudełko (black box) → 10 min wag - decyzja

XAI: algorytmiczne metody objaśnialności → Ruch QH1: 0,4 za króla h

Tłumaczenie na język ludzi → 10 wierszy długoterminowy atak na króla

Jak to osiągnąć?

HITL w szachach (Human-in-the-Loop)

Model

Człowiek (specjalista)

Propozycje AI

Feedback

Wynik końcowy

HUMAN THINKING

NON-HUMAN STRATEGIES

EXPERIENCE

ALGORITHMS

DATA ANALYSIS

Krótko o eXplainable AI

Jak szachiści mogą pomóc zrozumieć AI?

dr hab. Andrzej Śliódmok, prof. UJ Uniwersytet Jagielloński

"Chess helped me win the Nobel Prize" Hassabis

"so because of ... chess I started to think about thinking... and I started to improve my own thought processes... the computer play chess well I was intrigued and that got me into AI."

Problemy z interpretacją AI

Jak AI „myśli”?

Dane wejściowe (pozycja szachowa)

Sieć neuronowa (wzrosty wag)

Ocena pozycji (wartość ~0.7)

Symulacje (10^4 drzew gry)

Ruch: Qh1 (nieintuicyjny)

Nieintuicyjne cechy opisujące dane

Włoski na podstawie błędnych przesłanek = ryzyko błędu

Krótko o eXplainable AI

Roy Lopez (Hiszpańska) — po 3.805

SHAP (Shapley Additive Explanations)

- O ile każdy czepio podniósł lub obniżył wartość modelu dla konkretnego przybliżenia
- Wskazywał na najważniejsze czynniki (np. „grasowanie”) - Złaził wtedy między czepio, białe, czepio między, czepio czepio...
- Ma dwie funkcje użyteczności: **dominacja lokalna** (czepio walczyło = wygrał) i **spójność** (czepio czepio czepio czepio = wygrał, ale czepio czepio czepio czepio = przegrał)

Dlaczego potrzebujemy właśnie Was

- Ekspertka intuicja i plan: jak formułować objaśnienia. "Ask an expert"
- Domain knowledge: które objaśnienia są użyteczne, co realnie „działa” przy zegranie (i presji)

