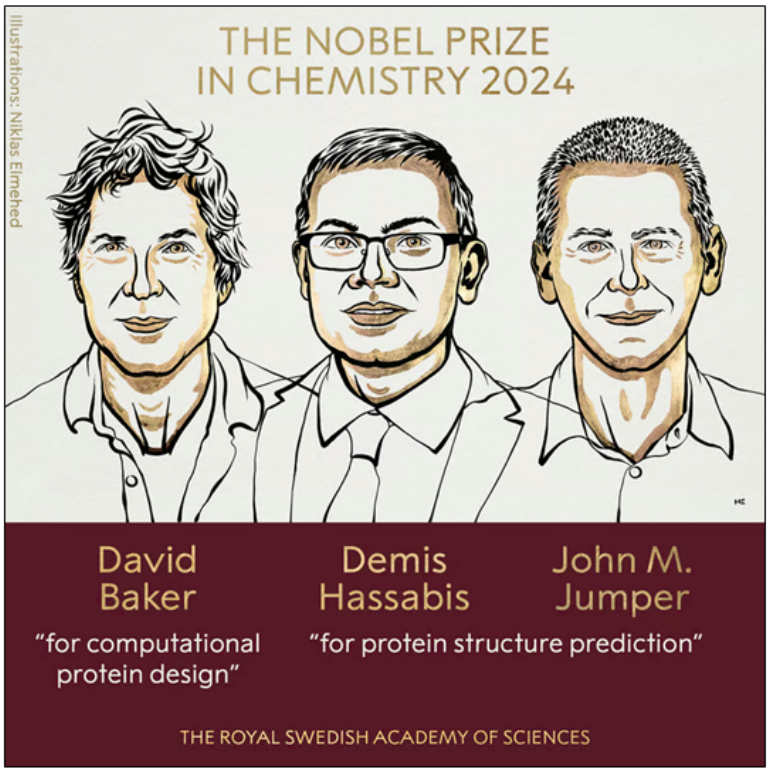
# Chess XAI study

**Weź udział w badaniu!** Zostaw swój adres e-mail poprzez formularz. Badanie będzie **anonimowe** – wyślemy Ci odpowiednie instrukcje, gdy rozpoczniemy eksperyment.

https://forms.gle/hEjpCTdZaka7sSsz8

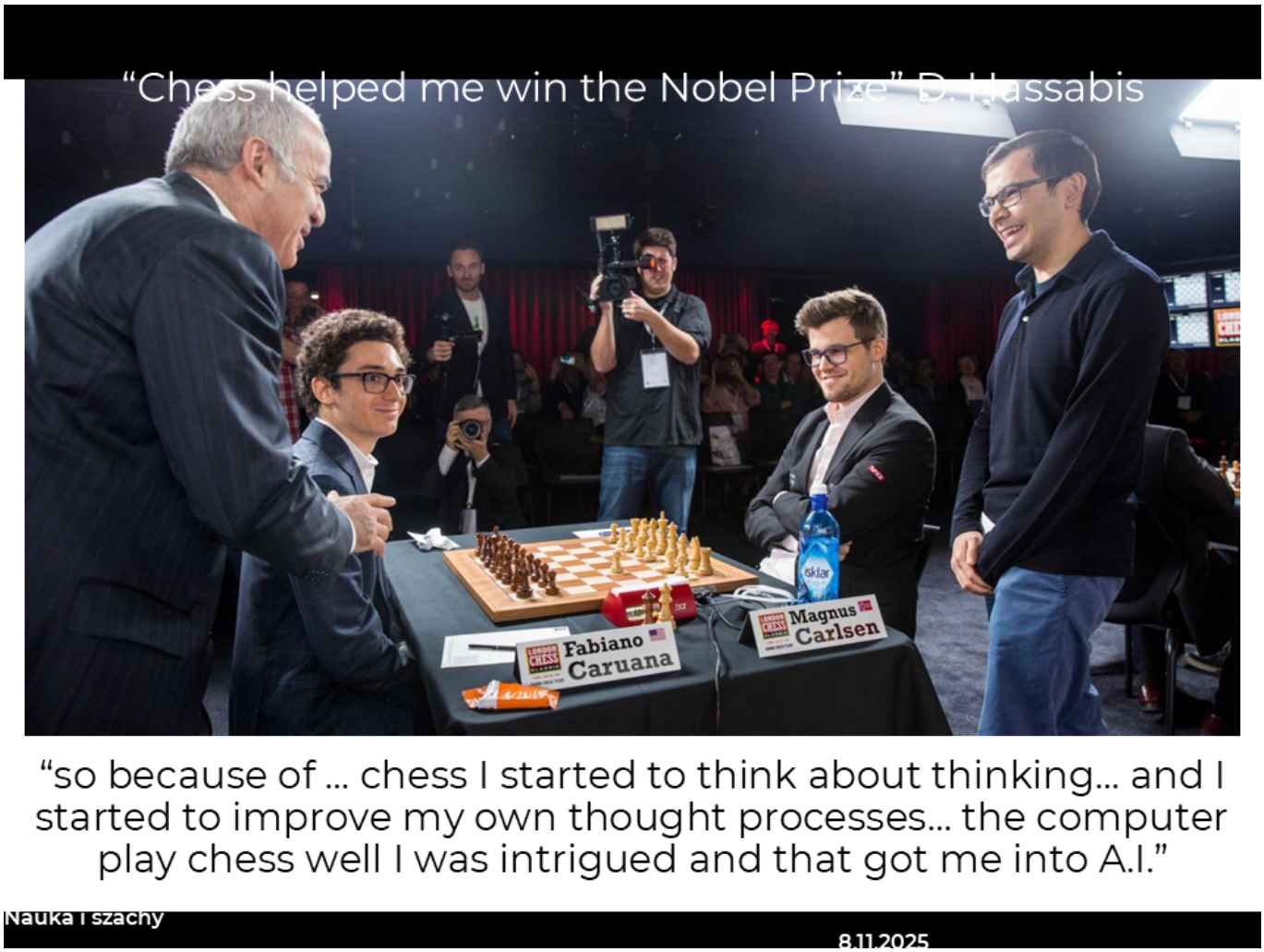
**Dziękujemy!**

# Rok temu: AlphaFold święty Graal biologii




Nauka i szachy

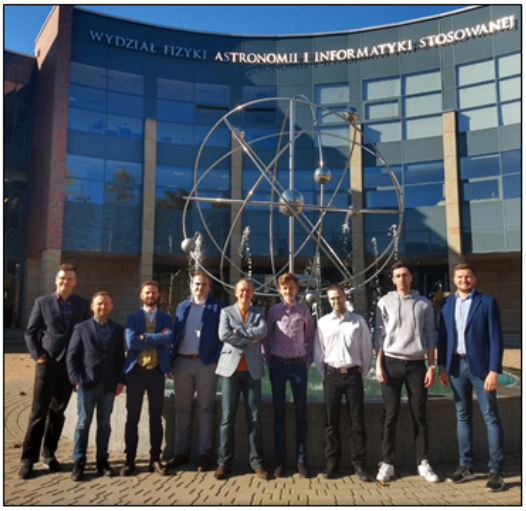
8.11.2025


“Chess helped me win the Nobel Prize” D. Hassabis
Fabiano Caruana
Magnus Carlsen
“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 A.I.”
Nauka i szachy
8.11.2025

# AI przyjazne i użyteczne dla ludzi

Zakład Humanocentrycznej Sztucznej Inteligencji
Uniwersytet Jagielloński

GEIST Research Group
We are GEIST. We dream big and work hard.




Maciej Szelążek

Maciej Mozolewski

***Zakład HCAI UJ oraz grupa GEIST kierowane przez prof. Grzegorza J. Nalepę***


Nauka i szachy

8.11.2025

# Problemy z interpretacją AI

## Jak AI „myśli”?

**Dane wejściowe**
(pozycja szachowa)

↓

**Sieć neuronowa**
(miliony wag)

↓

**Ocena pozycji**
(wartość: +0,7)

↓

**Symulacje**
($10^6$ drzew gry)

↓

**Qh1** **Ruch: Qh1**
(nieintuicyjny)

## Nieintuicyjne cechy opisujące dane

```
relative_castling_rights
relative_piece_connectivity
wk_fgh_12_bk_fgh_78
wk_not_12_bk_de_78
wk_not_12_bk_fgh_78
difference_king_safety_own_pieces
difference_castling_rights
.
.
```

## Wnioski na podstawie błędnych przesłanek = ryzyko błędu

(a) Husky classified as wolf

(b) Explanation

| | |
|---|---|
| Pytanie | *Husky czy wilk?* |
| Odpowiedź | *Husky* |
| Dlaczego? | *Wykryto śnieg* |

Nauka i szachy 8.11.2025

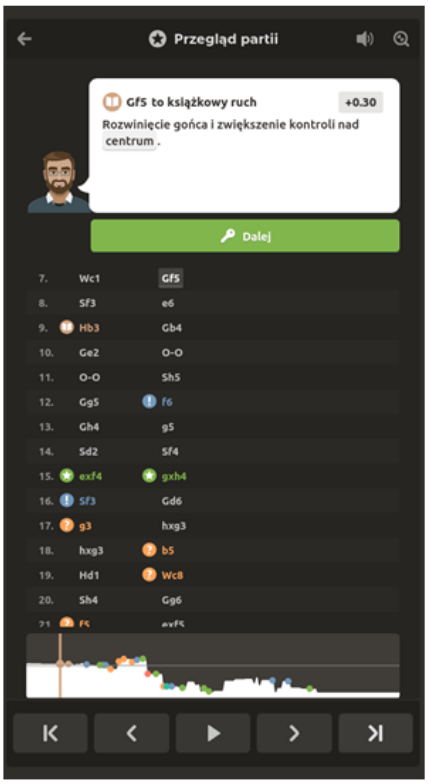
# Co chcemy zrobić?

## Jak objaśnić wnioskowanie?

**Czarne pudełko**
(black box)
→ 10 mln wag→ decyzja

**XAI: algorytmiczne metody objaśnialności**
→ Ruch Qh1: 0,4 za linię h

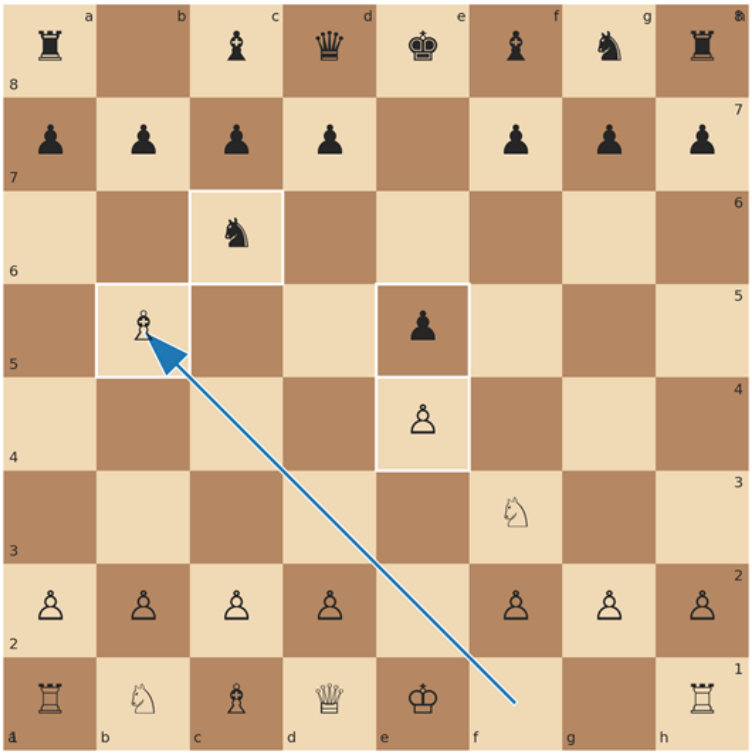
**Tłumaczenie na język ludzi**
→ AI widzi długoterminowy atak na króla




Nauka i szachy 8.11.2025

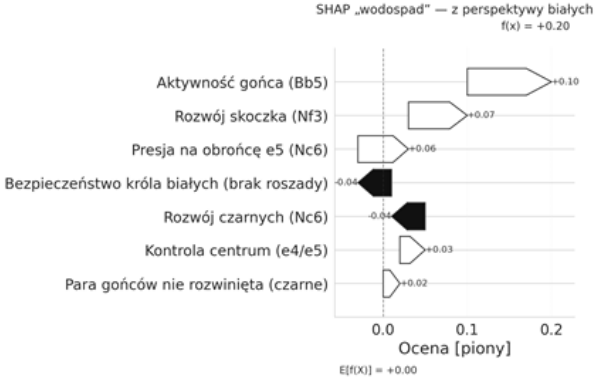
# Krótko o eXplainable AI

Ruy Lopez (Hiszpańska) — po 3.Bb5



## SHAP (Shapley Additive Explanations)

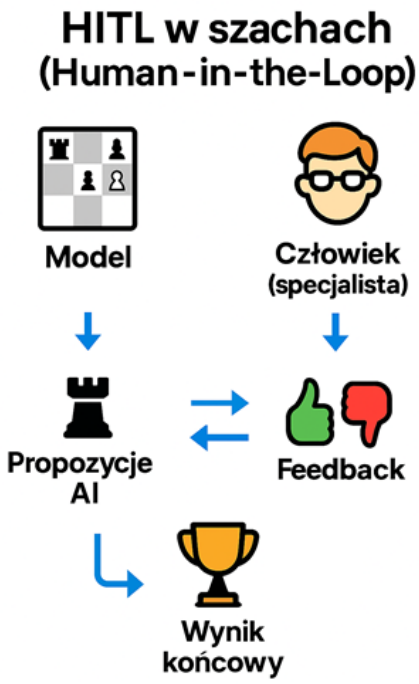
- **O ile każda cecha podnosi lub obniża** wynik modelu dla **konkretnego** przykładu.
- Wykorzystuje wartości **Shapleya** z teorii gier: „sprawiedliwie" dzieli wpływ między cechy, biorąc pod uwagę ich współdziałanie.
- Ma dwie kluczowe własności: **dokładność lokalną** (suma wkładów = wynik) i **spójność** (jeśli cecha pomaga bardziej w modelu, jej wkład nie spada).

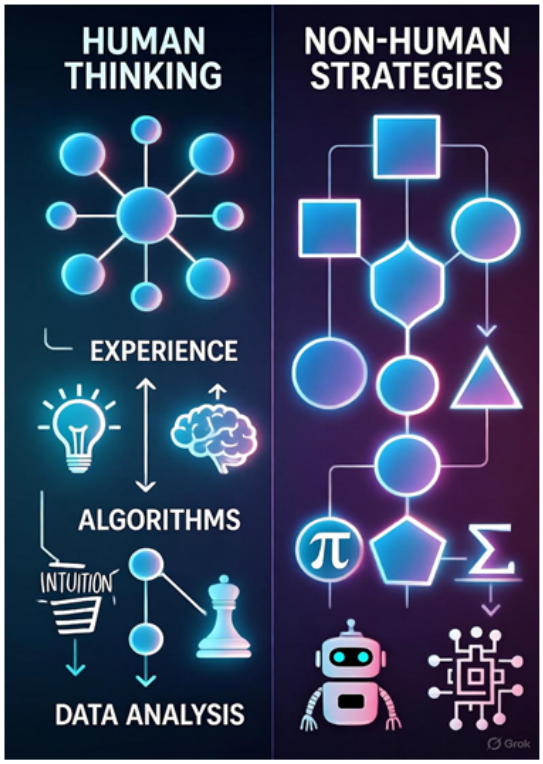



Nauka i szachy

8.11.2025

# Jak to osiągnąć?



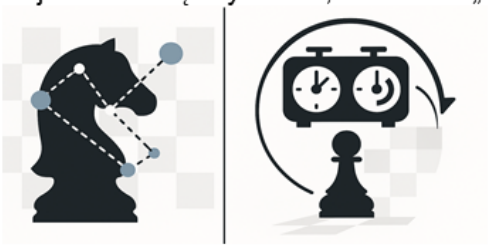

Nauka i szachy

8.11.2025

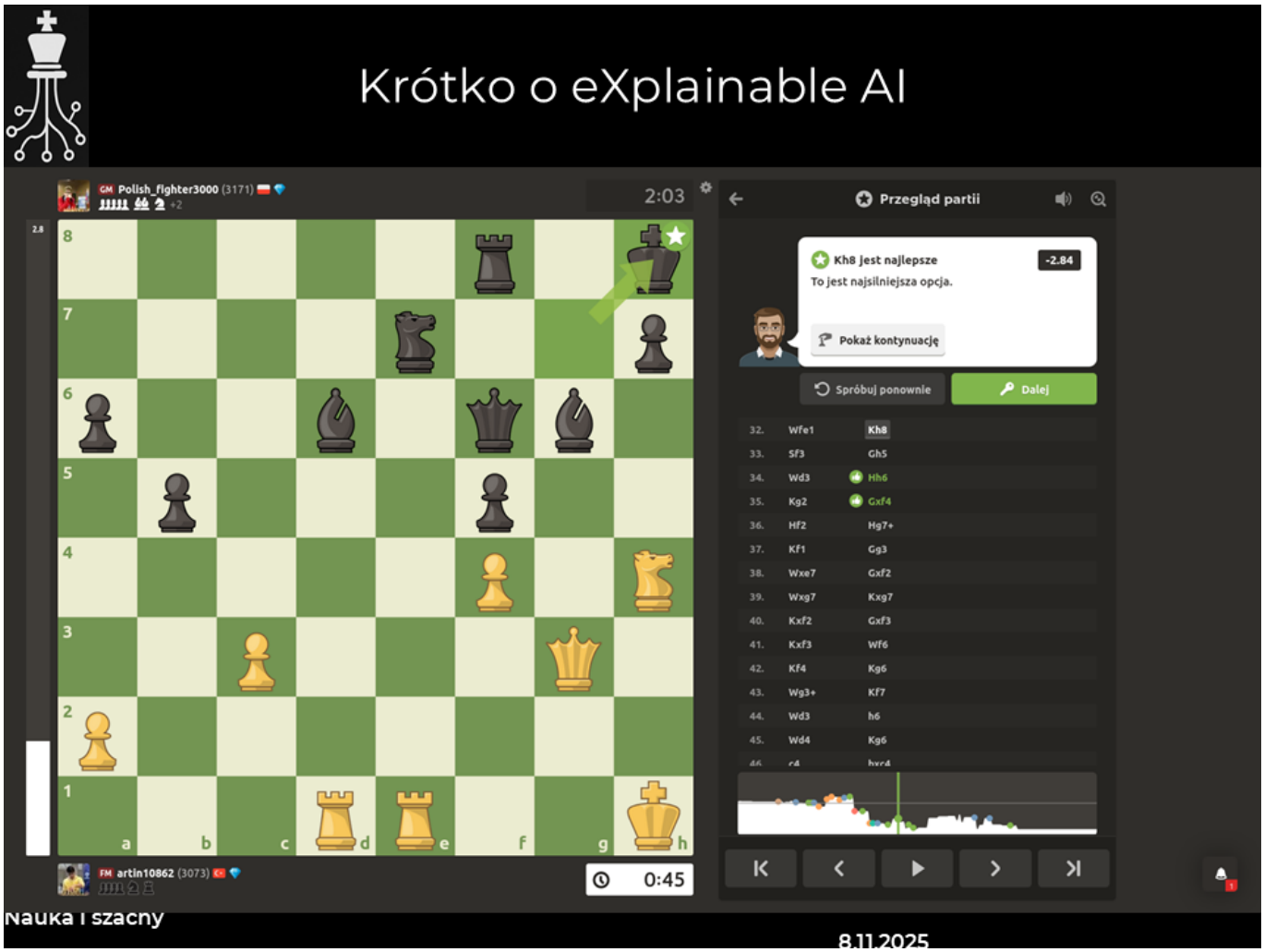
# Dlaczego potrzebujemy właśnie Was

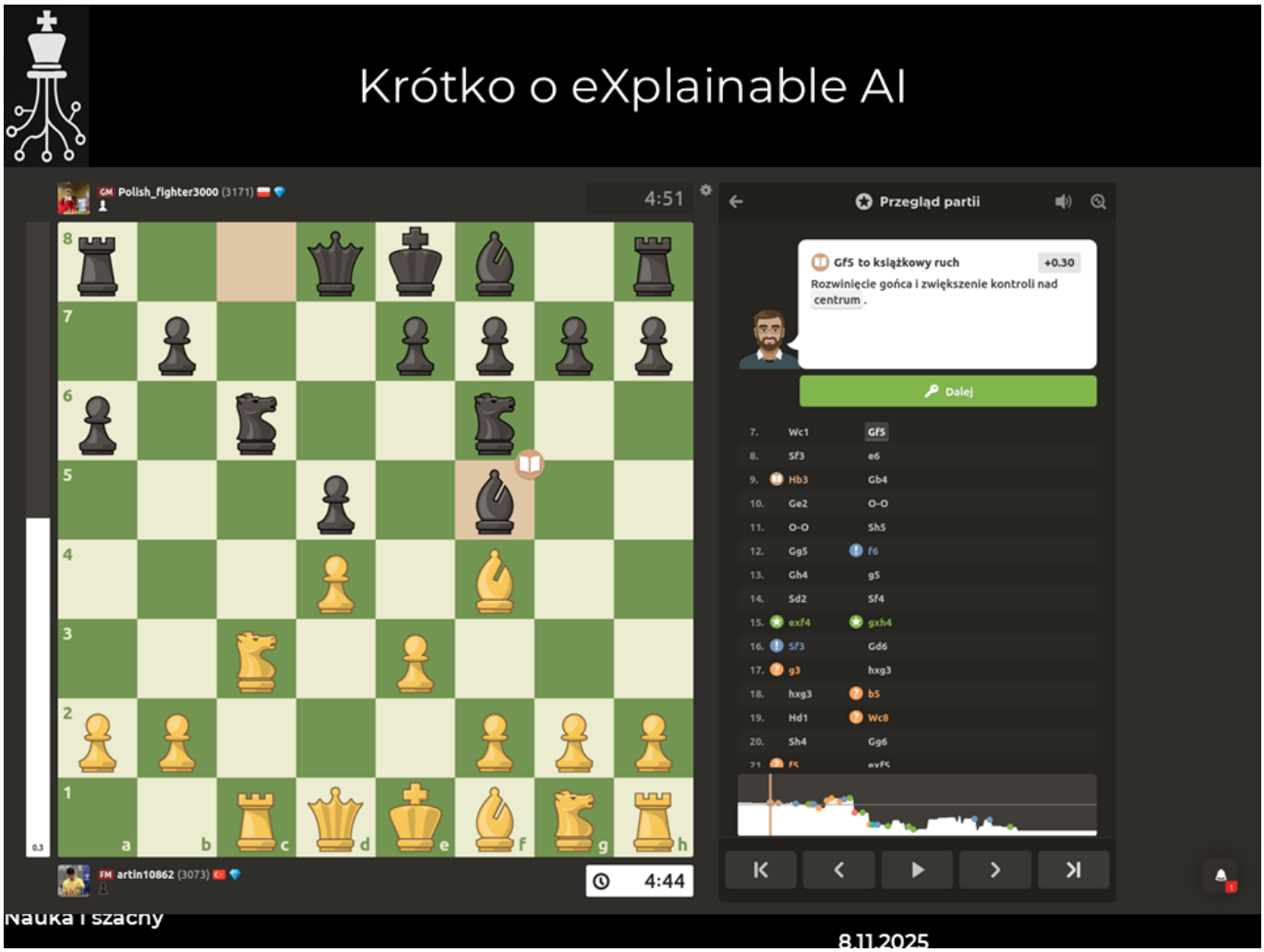
- Ekspercka intuicja i plan: jak formułować objaśnienia. “Ask an expert”


- Domain knowledge: które objaśnienia są użyteczne, co realnie „działa” przy zegarze i presji



Nauka i szachy 8.11.2025


Krótko o eXplainable AI
Polish_fighter3000 (3171)
2:03
Przegląd partii
Kh8 jest najlepsze
-2.84
To jest najsilniejsza opcja.
Pokaż kontynuację
Spróbuj ponownie
Dalej
32. Wfe1 Kh8
33. Sf3 Gh5
34. Wd3 Hh6
35. Kg2 Gxf4
36. Hf2 Hg7+
37. Kf1 Gg3
38. Wxe7 Gxf2
39. Wxg7 Kxg7
40. Kxf2 Gxf3
41. Kxf3 Wf6
42. Kf4 Kg6
43. Wg3+ Kf7
44. Wd3 h6
45. Wd4 Kg6
FM artin10862 (3073)
0:45
Nauka i szachy
8.11.2025


Krótko o eXplainable AI
GM Polish_fighter3000 (3171)
4:51
Przegląd partii
Gf5 to książkowy ruch
+0.30
Rozwinięcie gońca i zwiększenie kontroli nad centrum .
Dalej
7. Wc1 Gf5
8. Sf3 e6
9. Hb3 Gb4
10. Ge2 O-O
11. O-O Sh5
12. Gg5 f6
13. Gh4 g5
14. Sd2 Sf4
15. exf4 gxh4
16. Sf3 Gd6
17. g3 hxg3
18. hxg3 b5
19. Hd1 Wc8
20. Sh4 Gg6
FM artin10862 (3073)
4:44
Nauka i szachy
8.11.2025