Explanation in Artificial Intelligence: Insights from the Social Sciences

Prezentujący: Piotr Grabysz

Tim Miller

Profesor w School of Computing and Information Systems Uniwersytetu w Melbourne oraz dyrektor w Centre of AI and Digital Ethics

Zainteresowania:

- Interakcja i współpraca człowiek – AI
- Wyjaśnialne uczenie maszynowe (XAI)
- Podejmowanie decyzji w złożonych, wieloagentowych środowiskach

Źródło: https://people.eng.unimelb.edu.au/tmiller/

Co wyróżnia ten artykuł?

- Spojrzenie na to, jak ludzie przekazują między sobą wyjaśnienia
- Spojrzenie ze strony użytkownika korzystającego z wyjaśnień, nie inżyniera budującego model
- Czym jest dobre wyjaśnienie? - to pytanie wartościowe i niełatwe

Spis rzeczy

Spis rzeczy


Filozoficzne podłoże
Czym jest wyjaśnienie?
Społeczne atrybucje
Jak ludzie wyjaśniają zachowanie innych?
Procesy kognitywne
Jak ludzie wybierają i oceniają wyjaśnienia?
Społeczne wyjaśnienia
Jak ludzie przekazują wyjaśnienia?

Najważniejsze punkty:

1. Wyjaśnienia opierają się na kontrastach
2. Wyjaśnienia są wybiórcze
3. Prawdopodobieństwa nie mają większego znaczenia
4. Wyjaśnienia mają kontekst społeczny

Najważniejsze punkty:

1. Wyjaśnienia opierają się na kontrastach
2. Wyjaśnienia są wybiórcze
3. Prawdopodobieństwa nie mają większego znaczenia
4. Wyjaśnienia mają kontekst społeczny

Wyjaśnienie nie polega na wymienieniu przyczyn – ważny jest **kontekst**

Najważniejsze punkty:

1. **Wyjaśnienia opierają się na kontrastach**
2. Wyjaśnienia są wybiórcze
3. Prawdopodobieństwa nie mają większego znaczenia
4. Wyjaśnienia mają kontekst społeczny

Wyjaśnienia opierają się na przeciwieństwach

- Rozważania kontrfaktyczne: C jest przyczyną E, ponieważ gdyby nie wydarzyło się C, to nie mogłoby wydarzyć się E

Wyjaśnienia opierają się na przeciwieństwach

- Rozważania kontrfaktyczne: C jest przyczyną E, ponieważ gdyby nie wydarzyło się C, to nie mogłoby wydarzyć się E
- Wyjaśnienia są względne: dlaczego wydarzyło się P, skoro mogło wydarzyć się Q?

Wyjaśnienia opierają się na przeciwieństwach

- Rozważania kontrfaktyczne: C jest przyczyną E, ponieważ gdyby nie wydarzyło się C, to nie mogłoby wydarzyć się E
- Wyjaśnienia są względne: dlaczego wydarzyło się P, skoro mogło wydarzyć się Q?
- Większość badaczy jest zdania, że wszystkie pytania *dlaczego?* wymagają kontrastowego wyjaśnienia. Odpowiedź na pytanie *Dlaczego Ewa otworzyła drzwi?* zależy od (domyślnego) kontekstu:
  - Dlaczego Ewa otworzyła drzwi, zamiast zostawić je zamknięte?
  - Dlaczego Ewa otworzyła drzwi, zamiast otworzyć okno?
  - Dlaczego Ewa otworzyła drzwi, a nie Michał?

Wyjaśnienia opierają się na przeciwieństwach

- Rozważania kontrfaktyczne: C jest przyczyną E, ponieważ gdyby nie wydarzyło się C, to nie mogłoby wydarzyć się E
- Wyjaśnienia są względne: dlaczego wydarzyło się P, skoro mogło wydarzyć się Q?
- Większość badaczy jest zdania, że wszystkie pytania *dlaczego?* wymagają kontrastowego wyjaśnienia. Odpowiedź na pytanie *Dlaczego Ewa otworzyła drzwi?* zależy od (domyślnego) kontekstu:
  - Dlaczego Ewa otworzyła drzwi, zamiast zostawić je zamknięte?
  - Dlaczego Ewa otworzyła drzwi, zamiast otworzyć okno?
  - Dlaczego Ewa otworzyła drzwi, a nie Michał?

- Kontrastowe pytania bywają łatwiejsze do odpowiedzenia – nie trzeba znać wszystkich przyczyn

Model Arystotelesa typów wyjaśnień

1. Materialne
2. Formalne
3. Mechanicystyczne
4. Końcowe (funkcyjne)

Model Arystotelesa typów wyjaśnień

1. Materialne
2. Formalne
3. Mechanicystyczne
4. Końcowe (funkcyjne)

Dlaczego pióro zawiera atrament?

1. Pióro jest zrobione z substancji która zapobiega atrament od wylania się
2. Bo to pióro a pióra zawierają atrament
3. Ktoś napełnił je atramentem
4. Pióro służy do pisania, więc musi mieć atrament

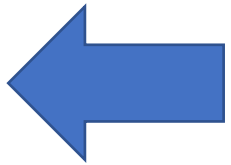
Bardziej złożony przykład

Jest pewien kwiat o nazwie holing. Holingi zazwyczaj mają związki bromu w swoich łodygach i zwykle wyginają się, gdy rosną. Naukowcy odkryli, że posiadanie związków bromu w łodygach jest tym, co zazwyczaj powoduje, że holingi wyginają się, gdy rosną.

Przez wyginanie się, pyłek holinga może ocierać się o sierść myszy polnych i w ten sposób rozprzestrzeniać się do terenów sąsiednich

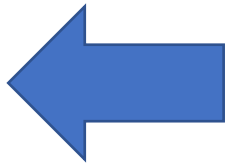
Bardziej złożony przykład

Jest pewien kwiat o nazwie holing. Holingi zazwyczaj mają związki bromu w swoich łodygach i zwykle wyginają się, gdy rosną. Naukowcy odkryli, że posiadanie związków bromu w łodygach jest tym, co zazwyczaj powoduje, że holingi wyginają się, gdy rosną.



Przyczyna mechaniczna

Przez wyginanie się, pyłek holinga może ocierać się o sierść myszy polnych i w ten sposób rozprzestrzeniać się do terenów sąsiednich



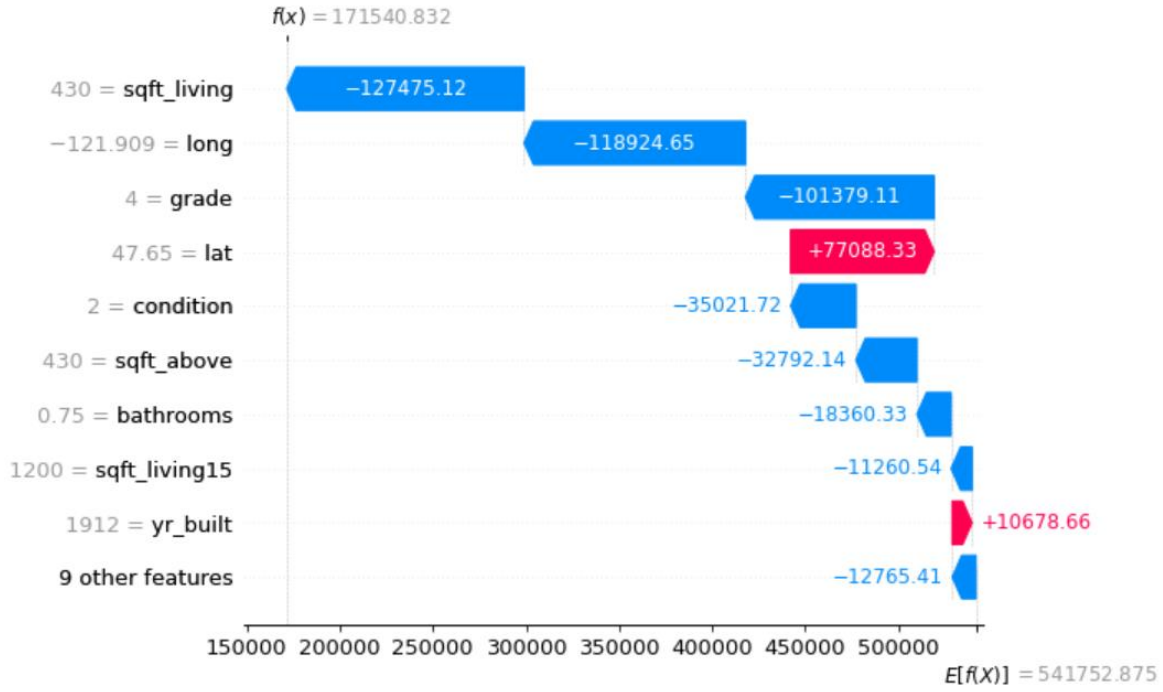
Przyczyna funkcjonalna

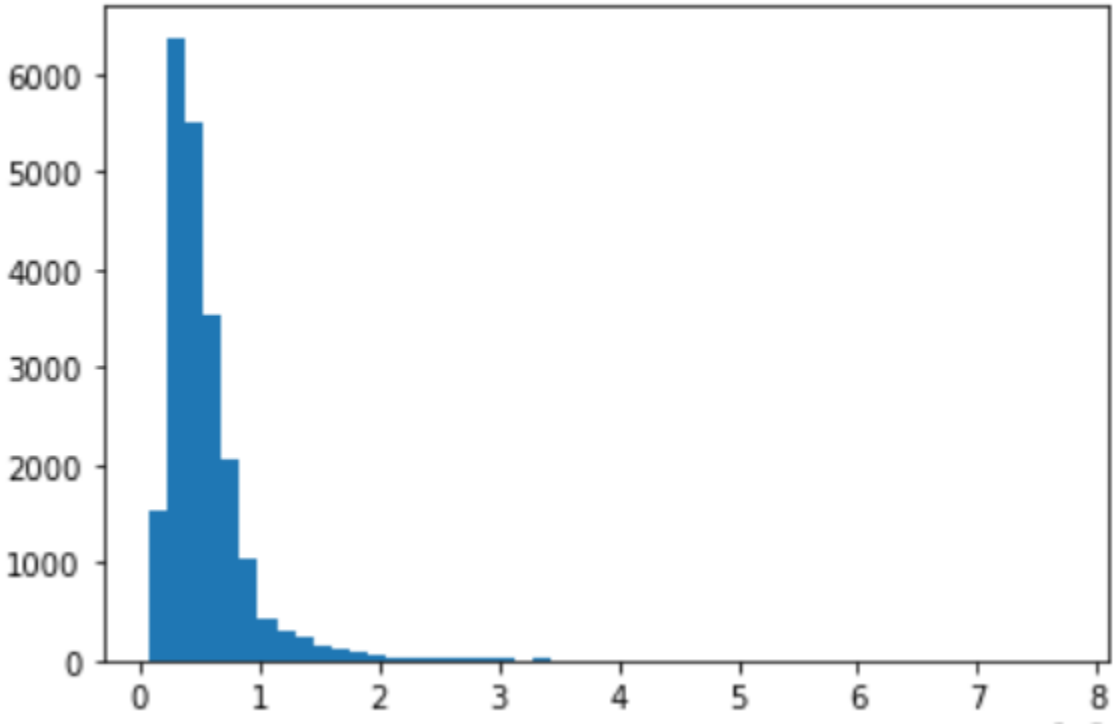
Pytanie: *Dlaczego holingi się wyginają?*

Dygresja – metryka jako *cecha* modelu

```
Prediction for the 'cheap' example: 171541$
The actual price: 80000$
The difference is 91541$
```

```
]: shap.plots.waterfall(shap_values[cheap_house_idx])
```





Prediction for the 'cheap' example: 171541$
The actual price: 80000$
The difference is 91541$

Najważniejsze punkty:

1. Wyjaśnienia opierają się na kontrastach
2. **Wyjaśnienia są wybiórcze**
3. Prawdopodobieństwa nie mają większego znaczenia
4. Wyjaśnienia mają kontekst społeczny

Dobór wyjaśnień

Śmierć osoby x

Lekarz	Prawnik	Konstruktor	Planista przestrzeni miejskich
To rozległy krwotok!	*To zaniedbanie kierowcy!*	*To wada w konstrukcji hamulca!*	*Krzewy rosną za wysoko na tym zakręcie!*

Żadne wyjaśnienie nie jest bardziej prawdziwe niż inne, ale szczególny **kontekst** pytania sprawia, że niektóre wyjaśnienia są bardziej **istotne** niż inne.

Cechy swoiste vs cechy zewnętrzne

- To to, co konstytuuje dany obiekt
- Pająki mają osiem nóg
- Ich nogi są przerażające

- Moi rodzice boją się pająków
- Arachnofobia może być dziedziczna

Najważniejsze punkty:

1. Wyjaśnienia opierają się na kontrastach
2. Wyjaśnienia są wybiórcze
3. **Prawdopodobieństwa nie mają większego znaczenia**
4. Wyjaśnienia mają kontekst społeczny

Cele i intencje

Zdarzenie: *Fred poszedł do restauracji*

Wyjaśnienia:

- *Fred był głodny* (oceniane jako rozsądne)
- *Fred miał przy sobie pieniądze* (niżej ocenianie)
- *Fred był głodny oraz miał przy sobie pieniądze* (najwyżej ocenianie)

„Najlepsze” wyjaśnienie jest najmniej prawdopodobne!

Chcę przyczyn!

Ludzie domagają się przyczyn w wyjaśnianiu zdarzeń a nie zależności statystycznych.

- Student: dlaczego dostałem tylko 50% z egzaminu?
- Nauczyciel: ponieważ większość studentów dostała około 50%
- *(lepsza odpowiedź)* Jaka jest przyczyna tego, że większość studentów dostała 50%?
- *(najlepsza odpowiedź)* Dlaczego ten konkretny student dostał 50%?

Wiarygodność a istotność

Pewne przekonanie psychologów

Grupa zastosowań
wymienić obserwacje które mogą być wytłumaczone za pomocą tego przekonania

Grupa wyjaśnień
wymienić obserwacje które mogą wyjaśnić to przekonanie

Grupa kontrolna
Nie robi nic

Przypisuje większe prawdopodobieństwo

Wiarygodność vs istotność

Pewne przekonanie psychologów

Grupa zastosowań
wymienić obserwacje które mogą być wytłumaczone za pomocą tego przekonania

Grupa wyjaśnień
wymienić obserwacje które mogą wyjaśnić to przekonanie

Grupa kontrolna
Nie robi nic

Przypisuje najniższą wartość!

Najważniejsze punkty:

1. Wyjaśnienia opierają się na kontrastach
2. Wyjaśnienia są wybiórcze
3. Prawdopodobieństwa nie mają większego znaczenia
4. **Wyjaśnienia mają kontekst społeczny**

Wyjaśnienie jako konwersacja

Istotność epistemiczna (poznawcza)

- Badani dostali policyjny raport na temat Grzesia, który został oskarżony o napaść w szkolnej bijatyce. W raporcie były informacje o samym Grzesiu, jak i okolicznościach bójki
- Badani mieli wytłumaczyć, dlaczego Grześ napadł na inną osobę, wiedząc że ich rozmówca:
  1. Zna informacje o Grzesiu
  2. Zna okoliczność bójki
  3. Nie ma żadnych informacji

Istotność epistemiczna (poznawcza)

- Badani dostali policyjny raport na temat Grzesia, który został oskarżony o napaść w szkolnej bijatyce. W raporcie były informacje o samym Grzesiu, jak i okolicznościach bójki
- Badani mieli wytłumaczyć, dlaczego Grześ napadł na inną osobę, wiedząc że ich rozmówca:
  1. Zna informacje o Grzesiu
  2. Zna okoliczność bójki
  3. Nie ma żadnych informacji
- Okazało się, że badani tworzyli wyjaśnienia zachowania Grzesia dopasowane do ich wyobrażenia na temat tego, co już wie ich rozmówca.

Efekt rozwodnienia wyjaśnień

Jaką średnią ocen osiągnie Dawid, jeśli:

- Grupa 1: uczy się 3 (C3) bądź 31 (C31) godzin tygodniowo
- Grupa 2: te same informacje, oraz mnóstwo nie powiązanych z uczeniem się (T3 oraz T31)

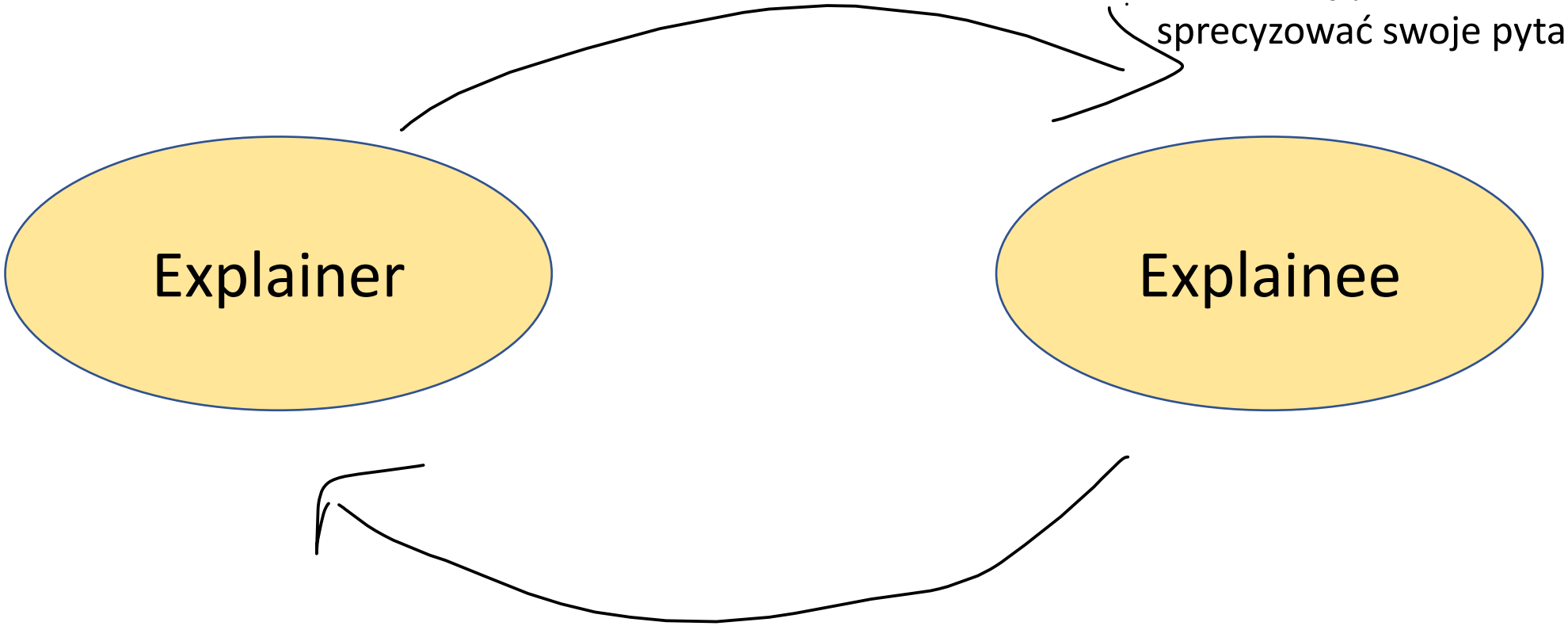
Efekt: Grupa T3 obstawiała wyższą średnią niż C3, grupa T31 niższą niż C31.

Wyjaśnienia są przekazywaniem wiedzy

- Jeśli coś dobrze rozumiemy, powinniśmy umieć odpowiedzieć na nowe, choć podobne pytania
- Tak działa np. nauka w szkole oraz na studiach

Wyjaśnienia jako spór

Odpowiedzi pozwalają osobie Explainee lepiej zrozumieć własne wątpliwości oraz sprecyzować swoje pytania



Pytania tworzą kontekst, przybliżają stan mentalny osoby Explainee

Dziękuję za uwagę!