W15. Etyka, bezpieczeństwo i przyszłość psychoinformatyki

Psychoinformatyka | Wykład 15

2026-11-04

Cele i zakres odpowiedzialności

- Oceniaj ryzyko dla osób, nie tylko trafność modelu.
- Rozróżniaj zgodę na badanie i podstawę RODO.
- Dokumentuj błędy i odtwarzalność analiz.

Etyka badań cyfrowych

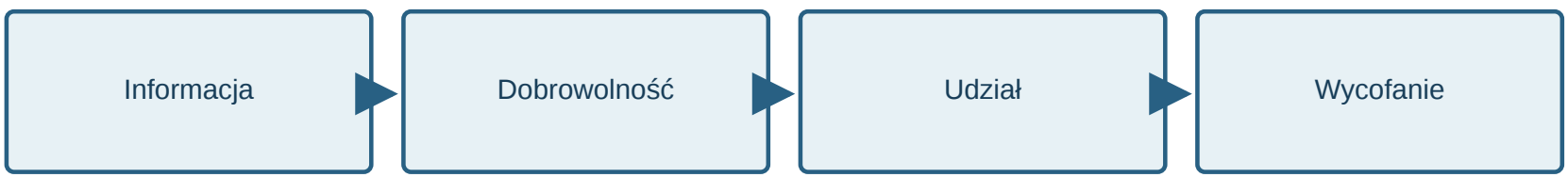
- Szanuj autonomię i godność uczestnika.
- Oceniaj ryzyko niezamierzonych szkód.
- Uwzględniaj asymetrię sił badacz–uczestnik.

Autonomia i dobrostan

- Ogranicz ingerencję do niezbędnego minimum.
- Umożliwiaj wycofanie udziału zgodnie z protokołem.
- Monitoruj niezamierzone skutki interwencji.

Świadoma zgoda na udział

- Przekaż cel, procedurę i konsekwencje uczestnictwa.
- Wyjaśnij dobrowolność i kanał kontaktu.
- Oddziel zgodę badawczą od podstawy prawnej RODO.



Rysunek 1: Świadoma zgoda na udział

Dane psychologiczne a identyfikacja

- Wypowiedzi i ślady interakcji mogą identyfikować.
- Wrażliwe informacje mogą wynikać z połączenia zbiorów.
- Dane pseudonimowe pozostają danymi osobowymi.

Zasady RODO

- Przetwarzaj zgodnie z prawem i przejrzyście.
- Minimalizuj cel, zakres i czas przechowywania.
- Zapewnij integralność i rozliczalność.

Szczególne kategorie danych

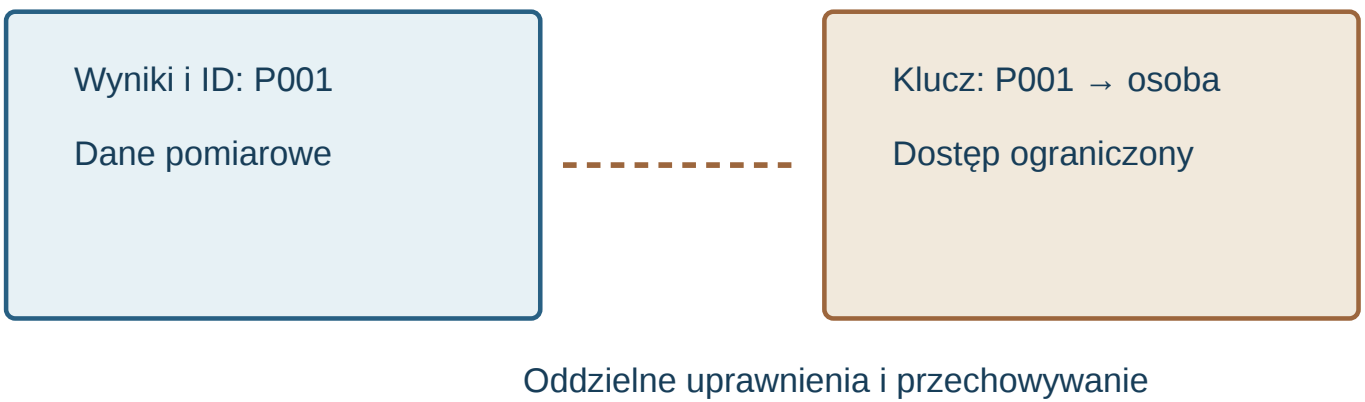
- Dane dotyczące zdrowia są szczególnie chronione.
- Dane biometryczne wymagają interpretacji definicji i celu użycia.
- Art. 9 określa warunki wyjątkowego przetwarzania.

Podstawa prawna i badania naukowe

- Ustal podstawę z art. 6 przed pobraniem danych.
- Dla danych szczególnych oceń również art. 9.
- Zgoda na udział nie musi być zgodą RODO.

Anonimizacja i pseudonimizacja

- Pseudonimizacja oddziela identyfikator od klucza.
- Anonimizacja wymaga oceny realnej możliwości identyfikacji.
- Nie gwarantuje jej samo usunięcie nazwiska.



Rysunek 2: Anonimizacja i pseudonimizacja

Ryzyko ponownej identyfikacji

- Zestawienie lokalizacji i czasu zwiększa ryzyko.
- Drobne podgrupy mogą być rozpoznawalne.
- Oceniaj wiedzę i środki potencjalnego przeciwnika.

Bezpieczne przechowywanie

- Szyfruj dane w spoczynku i transmisji.
- Ogranicz dostęp zasadą najmniejszych uprawnień.
- Określ retencję i sposób usuwania.

DPIA i analiza ryzyka

- Oceń wpływ przy prawdopodobnym wysokim ryzyku.
- Opisz cele, konieczność i środki ochronne.
- Zidentyfikuj ryzyko dla praw i wolności.

AI Act: ramy i harmonogram

- Rozporządzenie AI Act obowiązuje etapowo.
- Zakazy z rozdziału II stosuje się od lutego 2025.
- Wymogi wielu systemów wysokiego ryzyka mają późniejsze terminy.



Rysunek 3: AI Act: ramy i harmonogram

Aktualny tekst AI Act · art. 5 i art. 113.

Wnioskowanie o emocjach w edukacji

- Art. 5 ust. 1 lit. f zakazuje określonych zastosowań.
- Obejmuje wnioskowanie o emocjach w edukacji i pracy.
- Przewidziano wyjątki medyczne i bezpieczeństwa.

Art. 5 ust. 1 lit. f AI Act.

Studium: emocje studentów podczas zajęć

- Zakazane: wnioskowanie o emocjach studentów w zwykłej edukacji (art. 5 ust. 1 lit. f).
- Wysokiego ryzyka: AI do oceny efektów uczenia się (zał. III; obowiązki od 2.12.2027).
- Co do zasady dozwolone: badanie użyteczności bez inferencji emocji, z zachowaniem RODO.

5 minut przerwy

1. Czym anonimizacja różni się od pseudonimizacji?
2. Jak rozdzielić zgodę badawczą od RODO?
3. Czy zgoda studentów znosi zakaz z AI Act?
4. Jak interpretować lukę czułości modelu?

Systemy AI wysokiego ryzyka

- Klasyfikacja zależy od przeznaczenia systemu.
- Wymogi obejmują kontrolę jakości i dokumentację.
- Obowiązki uruchamiane są zgodnie z harmonogramem.

Przejrzystość i informowanie

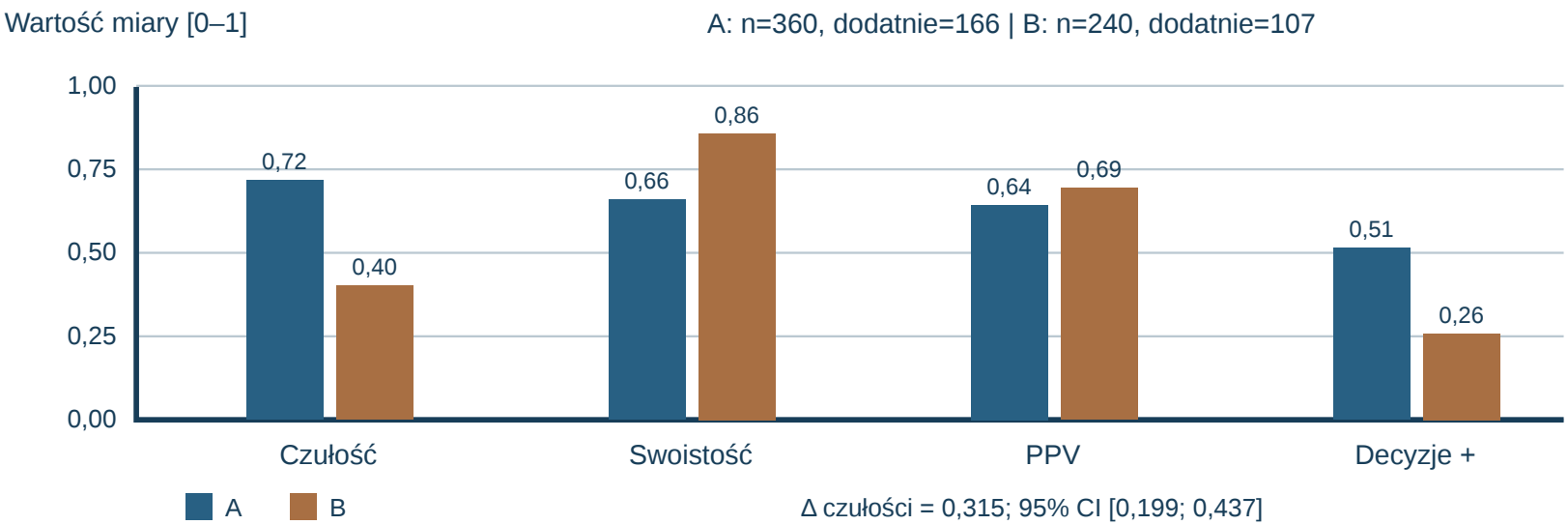
- Informuj o zakresie automatyzacji i konsekwencjach.
- Dokumentuj źródła danych i ograniczenia.
- Zapewnij procedurę zgłoszenia błędu.

Źródła stronniczości algorytmicznej

- Nierówny dobór osób zmienia reprezentatywność.
- Etykiety mogą odtwarzać błędy systemu opieki.
- Jakość pomiaru różni się między grupami.

Metryki sprawiedliwości

- Porównuj czułość i swoistość według podgrup.
- Raportuj liczebności i przedziały niepewności.
- Luka metryki nie dowodzi sama dyskryminacji.



Rysunek 4: Metryki sprawiedliwości

Wyjaśnialność modeli XAI

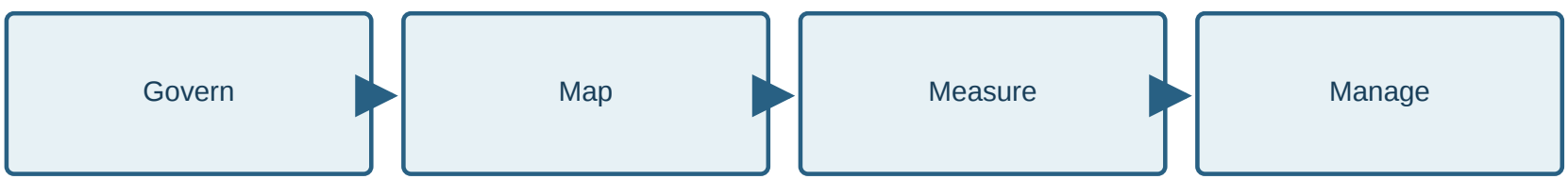
- Ważność cech opisuje zachowanie modelu.
- Wyjaśnienie lokalne nie ustala przyczynowości.
- Weryfikuj stabilność wyjaśnień.

Nadzór człowieka

- Osoba odpowiedzialna powinna rozumieć ograniczenia.
- Umożliwiaj korektę i zakwestionowanie decyzji.
- Projektuj bezpieczny tryb awaryjny.

Audyt i rejestr ryzyka

- Govern określa odpowiedzialność i procedury.
- Map identyfikuje zastosowanie i kontekst.
- Measure i Manage monitorują oraz ograniczają ryzyko.



Rysunek 5: Audyt i rejestr ryzyka

Python: sztuczne podgrupy

- Generujemy dwie grupy fikcyjnych obserwacji.
- Wynik modelu jest celowo przesunięty dla grupy B.
- Etykiety nie opisują rzeczywistych osób.

```
import numpy as np
import pandas as pd
from scipy.special import expit
rng = np.random.default_rng(42)
grupa = np.array(["A"]*360 + ["B"]*240)
cecha = rng.normal(size=len(grupa))
y = rng.binomial(1, expit(1.1*cecha-.15))
wynik = expit(1.2*cecha+rng.normal(0,.7,len(grupa))
              -.7*(grupa=="B"))
pred = (wynik >= .5).astype(int)
```

Python: miary jakości w grupach

- Wyznacz TN, FP, FN i TP oddzielnie.
- Policz czułość, swoistość, PPV i udział decyzji dodatnich.
- Zachowaj liczebności do interpretacji.

```
from sklearn.metrics import confusion_matrix,brier_score_loss
raport=[]
for g in ["A","B"]:
    maska=grupa==g
    tn,fp,fn,tp=confusion_matrix(y[maska],pred[maska],labels=[0,1]).ravel()
    raport.append({"grupa":g,"n":int(maska.sum()),
        "n_plus":int(tp+fn),"n_minus":int(tn+fp),
        "czulosc":tp/(tp+fn),"swoistosc":tn/(tn+fp),
        "ppv":tp/(tp+fp),"odsetek_decyzji_plus":(tp+fp)/maska.sum(),
        "brier":brier_score_loss(y[maska],wynik[maska]),
        "tp":int(tp),"fn":int(fn)})
df=pd.DataFrame(raport)
print(df.round(3).to_string(index=False))
```

Python: kalibracja w podgrupach

- Porównaj ryzyko przewidywane z częstością zdarzenia.
- Wyznacz pięć koszyków kwantylowych w każdej grupie.
- Oceniaj kalibrację wraz z liczebnością koszyków.

```
import matplotlib.pyplot as plt
kalibracja = {}
fig, ax = plt.subplots(figsize=(7.6, 3.7))
ax.plot([0, 1], [0, 1], "--", color="0.4", label="Idealna")
for g in ("A", "B"):
    ix = np.flatnonzero(grupa == g)
    koszyki = np.array_split(ix[np.argsort(wynik[ix])], 5)
    punkty = np.array([(wynik[b].mean(), y[b].mean(), len(b))
                       for b in koszyki])
    kalibracja[g] = punkty
    ax.plot(punkty[:, 0], punkty[:, 1], marker="o",
            label=f"{g} (n={len(ix)})")
    print(g, "liczebności koszyków:", punkty[:, 2].astype(int).tolist())
ax.set(xlabel="Średnia prognoza [0–1]",
       ylabel="Częstość wyniku [0–1]", xlim=(0, 1), ylim=(0, 1))
ax.legend(); ax.grid(alpha=.2)
plt.tight_layout(); plt.show()
```

Python: niepewność różnic

- Bootstrap przybliża niepewność luk metryk.
- Próbkuj oddzielnie wewnątrz grup.
- Przedział nie rozstrzyga kwestii prawnej dyskryminacji.

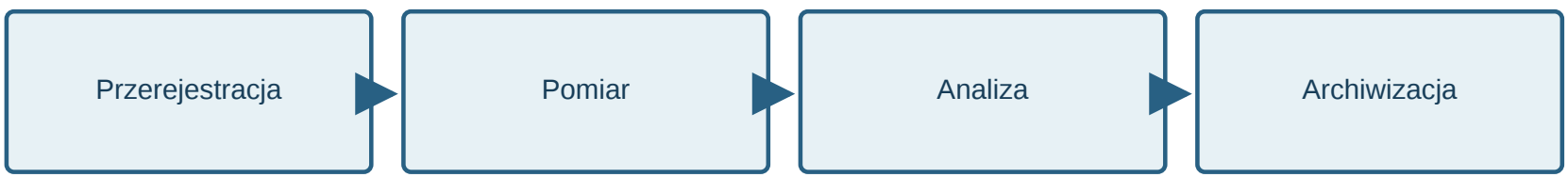
```
# Bootstrap różnicy czułości między grupami
rng_ci = np.random.default_rng(2026)
roznice = []
for _ in range(1000):
    sens = []
    for g in ["A","B"]:
        ix = np.flatnonzero(grupa==g)
        bs = rng_ci.choice(ix,size=len(ix),replace=True)
        mask = y[bs] == 1
        sens.append(np.mean(pred[bs][mask]) if np.any(mask) else np.nan)
    roznice.append(sens[0]-sens[1])
print("Różnica A–B, 95% bootstrap CI:",
      np.round(np.nanpercentile(roznice,[2.5,97.5]),3))
```

FAIR i otwarta nauka

- Findable, Accessible, Interoperable, Reusable opisują zarządzanie danymi.
- FAIR nie oznacza bezwarunkowej publikacji danych.
- Metadane i kod mogą być dostępne przy kontroli danych.

Reprodukowalność badań

- Przerejestruj hipotezy i kryteria wykluczeń.
- Zapisz wersje kodu, danych i zależności.
- Odtwarzaj wyniki na niezmienionym zestawie testowym.



Rysunek 6: Reprodukowalność badań

Literatura, akty prawne i standardy

- RODO (UE) 2016/679 i AI Act (UE) 2024/1689.
- NIST AI RMF, Model Cards i Datasheets.
- Wilkinson i in. (2016) – zasady FAIR.

RODO · AI Act – konsolidacja 2026 · NIST AI RMF · FAIR.

Model Cards (2019) · Datasheets for Datasets (2021).

Wnioski i synteza kursu

- Predykcja zachowania nie oznacza wyjaśnienia psychiki.
- Odpowiedzialność wymaga trafności, bezpieczeństwa i prawa.
- Sprawdź w pi_w15.ipynb różnice czułości między grupami.