UMCS
WYDZIAŁ FILOZOFII I SOCJOLOGII

Skrzywienie algorytmiczne (SI): teorie, typy i metody zwalczania

dr Kamil Filipek,
Instytut Socjologii
Uniwersytet Marii Curie-Skłodowskiej w Lublinie

Metodologiczne Inspiracje 2021: Badania ilościowe w naukach społecznych – wyzwania i problemy

Socjologia sztucznej inteligencji?

Socjologia sztucznej inteligencji

Socjologiczne zainteresowanie sztuczną inteligencją nie jest nowe:

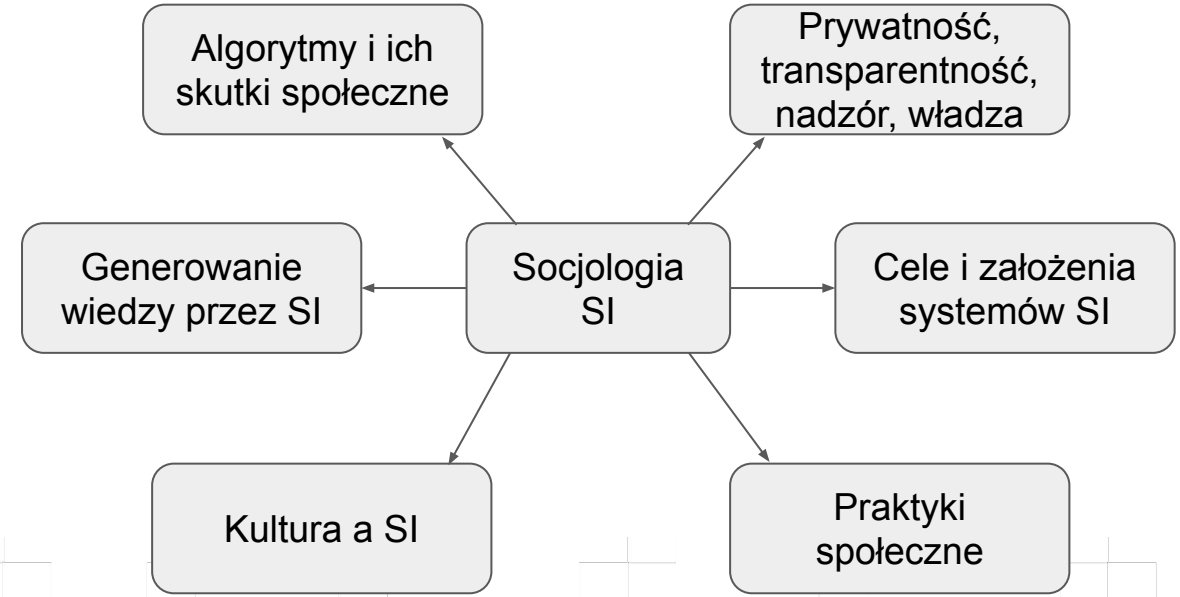
- **Woolgar, S. (1985)** → podziały “kognitywne-społeczne”, “człowiek-maszyna” ograniczają możliwości badawcze socjologii w obszarze SI.
- **Bloomfield, B. P. (1988)** → wiedza dot. SI jest wytwarzana przez ludzi, którzy są uwikłani w różne interesy społeczne, relacje, negocjacje itd (socjologia wiedzy). Wymiar społeczny wiedzy jest kluczowy w procesie rozwoju SI.
- **Schwartz, R. D. (1989)** → analiza fenomenu maszyn inteligentnych musi uwzględniać uwarunkowania społeczne, w których takie maszyny są implementowane.
- **Holton, R., Boyd, R. (2019)** → teoria socjologiczna jako ramy interpretacyjne współczesnych procesów rozwoju SI.
- **Liu, Z. (2021)** → trzy kategorie tekstów wokół AI: a) **naukowe** (SI ako nauka lub pole badań naukowych), b) **techniczne** (aspekty techniczne, aplikacje itd.), c) **kulturowe** (“perspektywa <<kulturowej SI>> postrzega rozwój SI jako zjawisko społeczne).

Socjologia sztucznej inteligencji

Schwartz, R. D. (1989). Artificial intelligence as a sociological phenomenon. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, 179-202.

“Socjologia sugeruje inny sposób rozumienia AI. Programy AI są pisane z **oczekiwaniem**, że **będą w stanie funkcjonować w sytuacjach społecznych**. Oznacza to badanie, **w jaki sposób działające programy łączą się z praktykami społecznymi**, absorbują je i przenikają do nich w sytuacjach, w których są wdrażane. Badanie takie wymaga nie tylko ocen technicznej adekwatności programu, ale także analizowania tego, **jak sami ludzie rozumieją program, co on robi, a także w jaki sposób organizacja i wymagania programu wpływają na relacje społeczne między ludźmi**. **Inteligencja nie tkwi w programie, ale jest rozpoznawana poprzez "działanie" programu w społecznym środowisku tworzonym przez ludzi**.”

Socjologia sztucznej inteligencji



Definicje

ALGORYTM, MODEL, SZTUCZNA INTELIGENCJA…

Algorytm jest procedurą (zaimplementowaną w kodzie), która uruchamiana jest na zbiorze danych. W praktyce, procedura ta obejmuje procesy **“uczenia się”** i **“dopasowania”** (ang. *fit*).

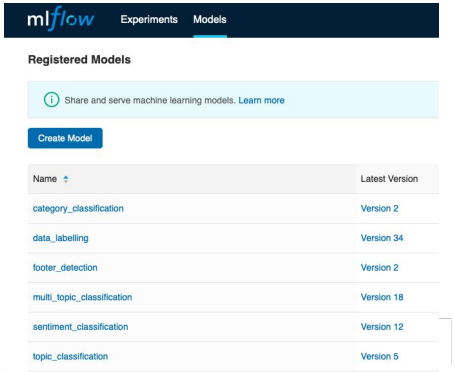
Model jest produktem/efektem działania algorytmu.

Sztuczna inteligencja to zbiór wytrenowanych modeli, które symulują działania wykonywane dotąd przez człowieka.

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
                                                    test_size=0.20, random_state=42)
X_train, X_val, Y_train, Y_val = train_test_split(X_train, Y_train,
                                                  test_size=0.20, random_state=42)
```

```
def _get_pipeline(self, dict_labels, top_words):
    model = tf.keras.Sequential([
        tf.keras.layers.Embedding(top_words, embedding_dim2),
        tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(100, dropout=0.5, recurrent_dropout=0.2,
                                                           return_sequences=True)),
        tf.keras.layers.LSTM(100, dropout=0.5, recurrent_dropout=0.2, return_sequences=True),
        tf.keras.layers.Dense(embedding_dim2, activation='relu'),
        tf.keras.layers.Dense(dict_labels, activation='sigmoid')
    ])
```

```
model.fit(X_train, Y_train, epochs=Epochs, validation_split=0.3, batch_size=batch_size,
          verbose=1, shuffle=True, callbacks=[callbacks]#, class_weight=weights
          )
return model
```



UMCS
UNIWERSYTET MARII CURIE-SKŁODOWSKIEJ

Definicje

SKRZYWIENIE CZY UPRZEDZENIE?

“Algorithmic bias” to **błąd** w systemach komputerowych, który **prowadzi do niepożądanych skutków.**

“Bias” → **uprzedzenie/a,** np. **uprzedzenia wobec grup ludzi** (PwC: *bias against groups of people*)

> *“Skłonność lub uprzedzenie do lub przeciwko jednej osobie lub grupie, szczególnie w sposób uważany za niesprawiedliwy.”* (Marshall 2013).
>
> *“Efekt uboczny działania algorytmu, a więc produkt uboczny świadomych i nieświadomych wyborów dokonywanych przez twórców i użytkowników algorytmów”.* (Baer, 2019).

W socjologii (naukach społecznych) “uprzedzenia” to również ang. **“prejudice”**. **Uprzedzenia zwykle kojarzą się pejoratywnie**, tj. wywołują negatywne skutki społeczne. Tymczasem “bias” może prowadzić do **negatywnych**, ale i **pozytywnych** skutków społecznych.

Wydaje się zatem, że mniej problematyczne jest używanie pojęcia “skrzywienie algorytmiczne”.

Dlaczego sztuczna inteligencja działa nieprawidłowo?

Skrzywiona sztuczna inteligencja

KIEDY SI STAJE SIĘ SKRZYWIONA?

...kiedy podejmuje decyzje, które są **błędne z punktu widzenia skutków**, do których prowadzi (rekruter Amazon'a, chatbot Microsoft'u, algorytm predykcji pomocy medycznej w USA, full-self-driving Tesla itd.)

CZY JESTEŚMY W STANIE PRZEWIDZIEĆ SKUTKI DZIAŁANIA SI?

...nie zawsze, ale bez wiedzy na temat **celu działania modelu (!)** nie jesteśmy w stanie zidentyfikować jego skrzywienia.

Skrzywiona sztuczna inteligencja

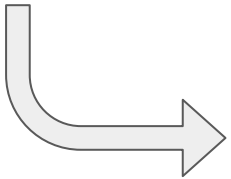
PRZYCZYNY SKRZYWIENIA ALGORYTMICZNEGO:

1. **Błąd doboru próby** - na etapie zbierania danych
2. **Błąd wykluczenia** - na etapie selekcji danych treningowych
3. **Błąd pomiaru** - dane treningowe nie odpowiadają danym rzeczywistym
4. **Błąd niespójności** - ruchome granice semantyczne między kategoriami (labelkami)
5. **Błąd potwierdzenia** - negatywny wpływ prekoncepcji (wiedzy) na postrzeganie danych
6. **Błąd asocjacji** - dane treningowe wzmacniają uprzedzenia rasowe, płciowe, etniczne itd.

Skrzywiona sztuczna inteligencja

Przykład skrzywienia w NLP:

“The doctor gave instructions to the nurse before she left.”

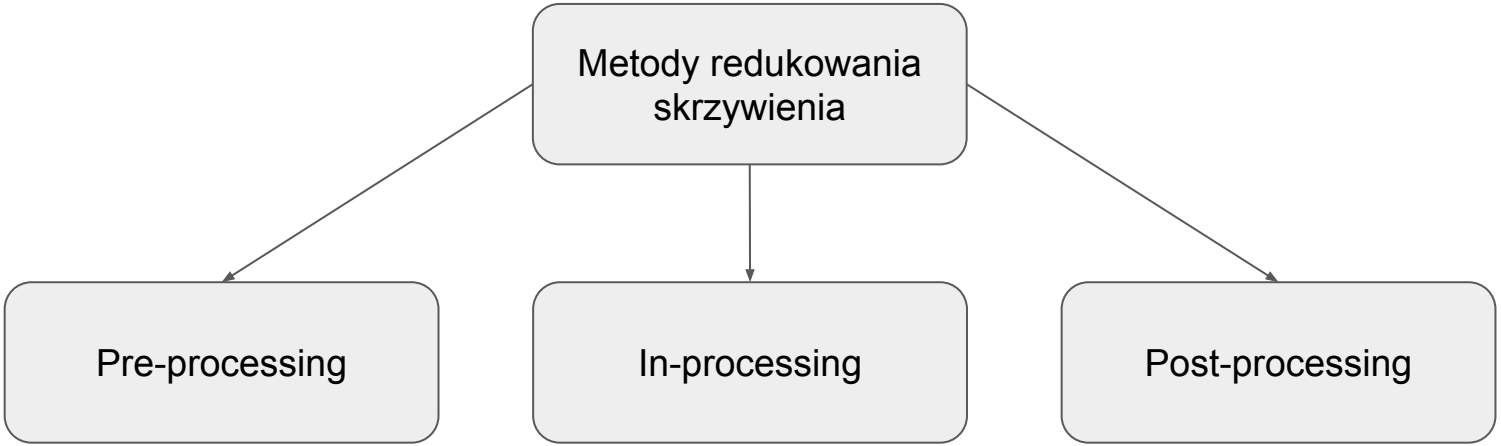


he

Przykład skrzywionego zakorzenienia na poziomie modeli językowych. Zawód - płeć

Jak walczyć ze skrzywieniem SI?

Klasyfikacja metod



https://aif360.readthedocs.io/en/latest/modules/algorithms.html#id17

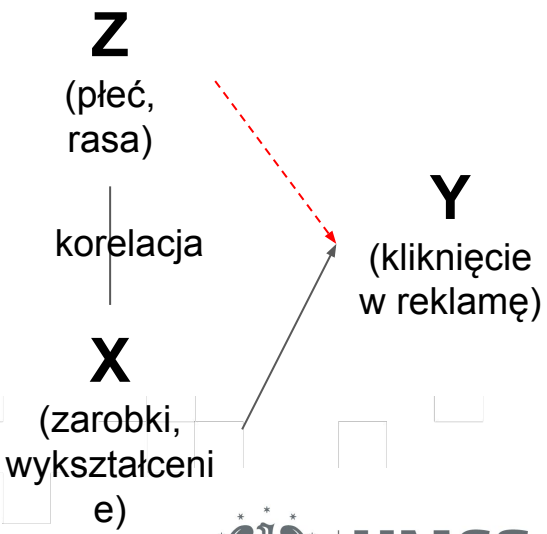
1. Współzawodniczące eliminowanie skrzywienia (pre)

Adversarial de-biasing - trenowanie przeciwstawne w celu usunięcia skrzywienia powstałego z ukrytych (latent) reprezentacji w wyuczonym modelu (model generatywny).

W dużym uproszczeniu, **dane treningowe wzbogaca się próbkami nieznacznie zmodyfikowanymi, w celu osłabienia ukrytych zależności, które występują w danych oryginalnych.**

Poprawna klasyfikacja jest dalej celem działania modelu (zmienna Y pozostaje bez zmian), ale X1,2,...n zostają zmodyfikowane po to, by usuwać latentne zależności wywołujące skrzywienie.

W praktyce, budujemy dwa modele (sieci): 1. X → Y, 2. Y → Z.



* poprzez próbkowanie z takiego modelu jesteśmy w stanie wygenerować nowe dane.

UMCS
UNIWERSYTET MARII CURIE-SKŁODOWSKIEJ

1. Współzawodniczące eliminowanie skrzywienia

$$\min \left[\sum_{(x,y)\in X} L_Y(f(g(x)), y) + \sum_{(x,z)\in S} L_Z(a(J_\lambda(g(x))), z) \right]$$

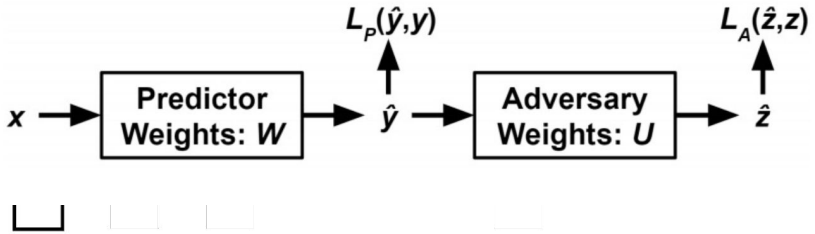
Matematyczny cel trenowania obiektu: **minimalizacja dwóch funkcji straty Ly** (*normal loss* Ly(f(g(X)), Y)) **i Lz** (*cross entropy classification loss* Lz(a(g(S)), Z).

Gdyby Lz zostało tak jak jest wyżej, wówczas zostałoby zachęcone do przewidywania Z. Potrzebne jest więc odwrócenie zależności Lz(a(Jλ(g(S))), Z); Jλ - to funkcja, która wskazuje kierunek najszybszego spadku (ang. *negative gradient*).

1. Współzawodniczące eliminowanie skrzywienia

Generative Adversarial Networks (GAN) - dwie niezależne sieci neuronowe, pierwsza zwana **dyskryminatorem**, którą uczymy rozpoznawać obrazy, i druga, która uczy się generować obrazy (**generatora**) grają w swoistą grę.

W trakcie procesu uczenia obydwa modele podnoszą swoje umiejętności. Generator generuje coraz lepsze zdjęcia, a dyskryminator umie coraz lepiej rozpoznawać zdjęcia nieprawdziwe i prawdziwe. **W momencie gdy generator zacznie tworzyć tak realistyczne zdjęcia, że dyskryminator nie będzie w stanie odróżnić ich od prawdziwych, model generatywny zostanie w pełni wytrenowany.** Dzięki temu będzie on w stanie generować na żądanie zdjęcia, które będą bardzo realistyczne.

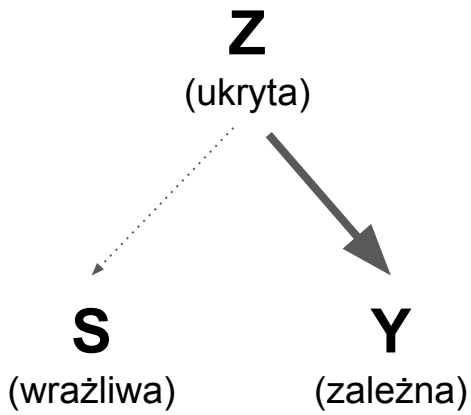


2. Wariacyjne autokodowanie (in)

Variational "fair" autoencoders - model, który uczy się parametryzacji z pewnej utajonej reprezentacji danych wejściowych (model generatywny).

Chodzi o **identyfikację reprezentacji (przestrzeni) ukrytej Z**, która jest maksymalnie informuje o obserwowanej zmiennej losowej Y (np. etykieta klasy), a jednocześnie minimalnie informuje o wrażliwej lub uciążliwej zmiennej S.

Autoencoder - to pewna funkcja kodująca, a następnie dekodująca dane, przy określonej stracie (ang. *loss function*).



UMCS
UNIWERSYTET MARII CURIE-SKŁODOWSKIEJ

2. Wariacyjne autokodowanie

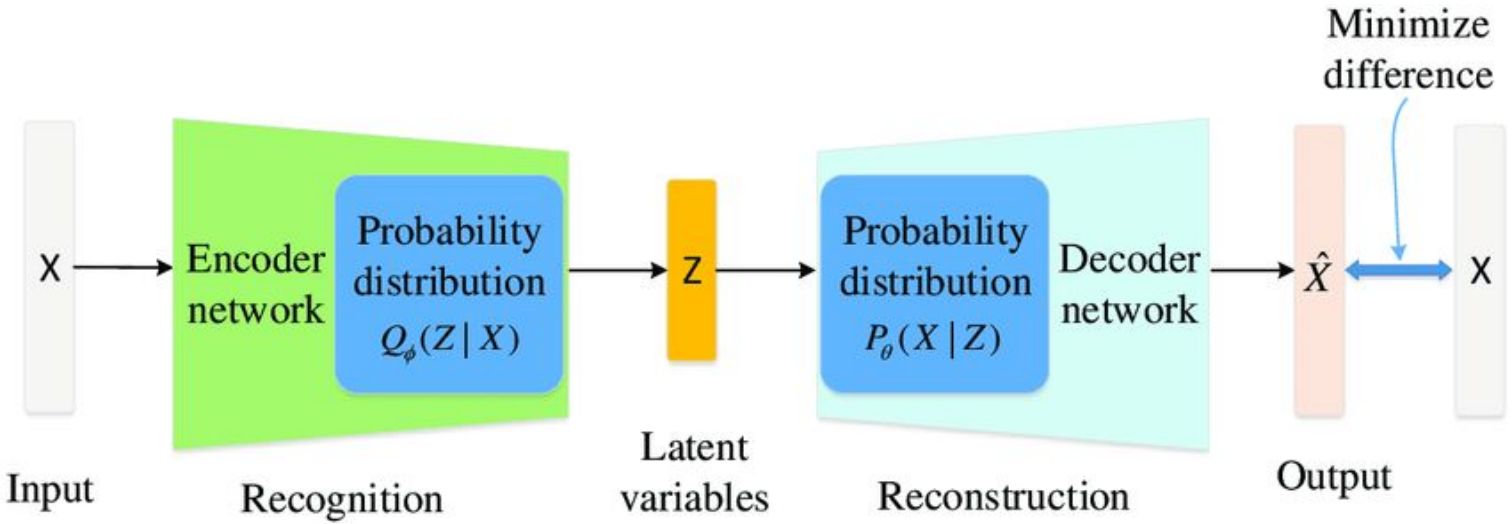
Autokoder wariacyjny

1. Otrzymuje dane wejściowe jako dane wielowymiarowe
2. Kompresuje jest do przestrzeni utajonej (kodowanie)
3. Rekonstruuje dane (dekodowanie)

CEL: minimalizacja straty funkcji rekonstrukcji danych

W praktyce, chodzi o poznanie parametrów rozkładu zmiennej ukrytej, po to by nadać wagi wybranym danym (data points) podczas treningu.

2. Wariacyjne autokodowanie



3. Wyrównywanie szans (post)

Equalized odds - inaczej równość warunkowej procedury dokładności, to jedna z metod zapewniania równości w uczeniu maszynowym.

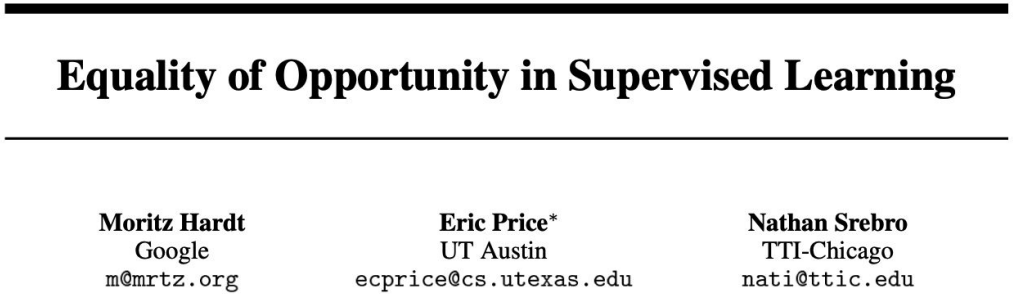
Klasyfikator spełnia tę definicję, jeśli **badani w grupie chronionej i niechronionej mają równy współczynnik klasyfikacji prawdziwie pozytywnych** (true positives) i **równy współczynnik klasyfikacji fałszywie pozytywnych** co wyraża wzór:

$$P(R = +|Y = y, A = a) = P(R = +|Y = y, A = b) \quad y \in \{+, -\} \quad \forall a, b \in A$$

A - rasa studenta, Y - czy student spełnia warunki studiowania, R - decyzja o przyjęciu do szkoły. Jeśli studenci biali i czarni, o takich samych wynikach będących podstawą przyjęcia na studia, są przyjmowani w takich samych proporcjach, klasyfikator spełnia swoje zadanie.

3. Wyrównywanie szans

Najważniejszy tekst dotyczący tej metody:



Equality of Opportunity in Supervised Learning

Moritz Hardt
Google
m@mrtz.org

Eric Price*
UT Austin
ecprice@cs.utexas.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu

https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

UMCS
UNIWERSYTET MARII CURIE-SKŁODOWSKIEJ

Wybór metody

Wybór metody

Wybór metody redukowania skrzywienia mocno powiązany jest z teorią sprawiedliwości (ang. *fairness*), z którą utożsamiają się badacze SI:

1. **Sprawiedliwość dystrybutywna -** ważne jest zapewnienie równości na etapie uczenia modelu, w taki sposób, że grupy mniejszościowe są dowartościowane (Rawls, 1971).
2. **Sprawiedliwość proceduralna -** ważny jest wgląd w działanie algorytmu w celu zapewnienie jego spójności i przejrzystości. Chodzi o monitorowanie procedury działania algorytmu (Hadzi & Rojo, 2019).
3. **Sprawiedliwość naprawcza/retrybutywna** - SI może działać niesprawiedliwie, zaakceptujmy to, ale zastanówmy się jak zrekompensować ofierze (**naprawcza**) szkody, które wywołuje SI (Hadzi & Rojo, 2019). W innym przypadku, możemy zrekompensować złe działanie algorytmy karząc sprawcę (np. organizację) (**retrybutywna**). (De Diego Carreras, 2020).

Wybór metody

Warto jednak zaznaczyć, że wspomniane teorie **nie wyczerpują możliwości "kontekstualizacji" rozwoju sztucznej inteligencji**.

Można zaryzykować stwierdzenie, że ciągłe powoływanie się w/w teorie sprawiedliwości pokazuje, jak bardzo **"niedoteoretyzowany" jest obszar rozwoju sztucznej inteligencji**.

Wyzwania dla socjologa

Wyzwania

- Jakie **teorie / podejścia teoretyczne** na gruncie nauk społecznych można **wykorzystać do identyfikacji niesprawiedliwości w danych**, które służą trenowaniu modeli SI?
- Czy jesteśmy w stanie dokonać swoistej “**translacji” teorii na** konkretny **algorytm** zawarty w kodzie budującym model?
- W jaki sposób **projektować modele by zachęcać ludzi do korzystania z nich** (w efekcie dostarczania danych do retreningu)?
- Czy istniejące **miary skuteczności modeli** pokrywają się z **miarami efektywności na poziomie społecznym**? (Accuracy modele nie oznacza skuteczności na poziomie realnych decyzji podejmowanych w oparciu o model)

Bibliografia

- Bloomfield, B. P. (1988). Expert systems and human knowledge: a view from the sociology of science. *AI & SOCIETY*, *2*(1), 17-29.
- De Diego Carreras, A. (2020). THE MORAL (UN) INTELLIGENCE PROBLEM OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE: A COMPARATIVE ANALYSIS UNDER DIFFERENT THEORIES OF PUNISHMENT. *UCLA Journal of Law & Technology*, *25*(1).
- Hadzi, A., & Roio, D. (2019). Restorative Justice in Artificial Intelligence Crimes. *spheres: Journal for Digital Cultures*, (5), 1-18.
- Holton, R., & Boyd, R. (2021). 'Where are the people? What are they doing? Why are they doing it?'(Mindell) Situating artificial intelligence within a socio-technical framework. *Journal of Sociology*, *57*(2), 179-195.
- Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass*, *15*(3), e12851.
- Marshall, D. (2013). "Recognizing your unconscious bias," *Business Matters*, www.bmmagazine.co.uk/in-business/recognising-unconscious-bias/, October 22, 2013.
- Rawls, J. (1971), A Theory of Justice, Harvard University Press,
- Schwartz, R. D. (1989). Artificial intelligence as a sociological phenomenon. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, 179-202.
- Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology*, *19*(4), 557-572.

UMCS
UNIWERSYTET MARII CURIE-SKŁODOWSKIEJ