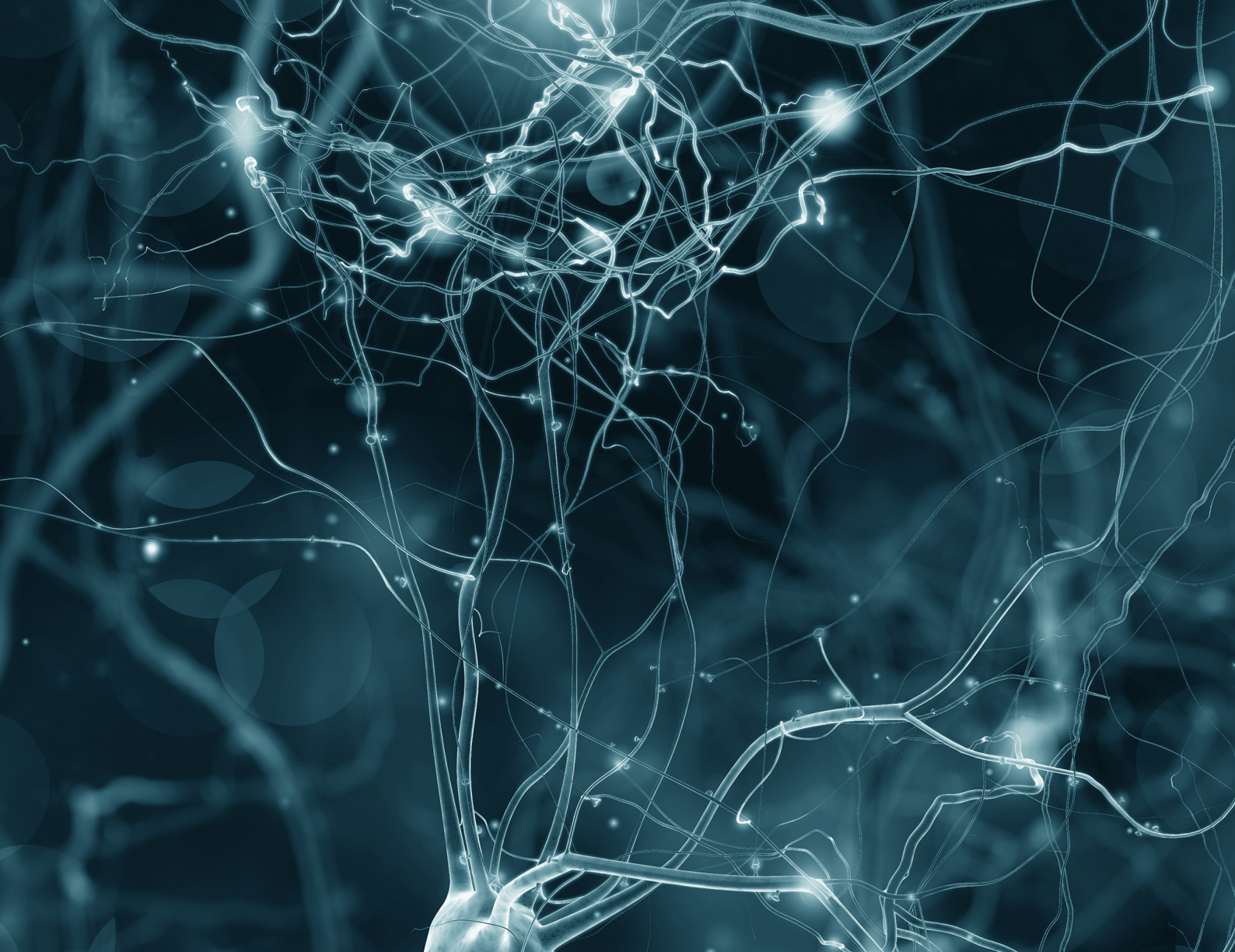




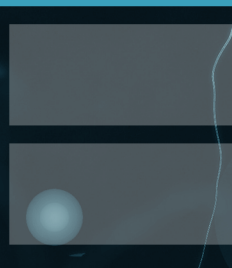
**AGiDOT**

# Równanie płci: Co mówią nam modele językowe?

W jaki sposób modele sztucznej inteligencji  
mogą wzmacniać stereotypy?

Patronat medialny:

Sponsorzy:



# Zapowiedź raportu

## Abstrakt

W jaki sposób modele sztucznej inteligencji mogą wzmacniać stereotypy?

W tym raporcie przyglądamy się, jak sztuczna inteligencja (AI) przetwarza i rozumie język w kontekście płci. Modele językowe, takie jak te stosowane w wyszukiwarkach czy chatbotach, uczą się na podstawie ogromnych zbiorów danych z internetu. Problem polega na tym, że w tych danych mogą być ukryte uprzedzenia i stereotypy, które AI może nieświadomie wzmacniać.

Jednym z narzędzi, których używamy do badania tego zjawiska, są tzw. **embeddingi**. To sposób, w jaki AI reprezentuje słowa jako liczby. Dzięki temu model potrafi rozpoznawać związki między nimi. Analizując, jak model językowy "widzi" różne słowa związane z płcią, możemy sprawdzić, czy istnieją różnice w skojarzeniach np. między "mężczyzną" a "kobietą" w kontekście pracy.

## Wprowadzenie

### Cel raportu

Celem naszego raportu jest pokazanie, jak sztuczna inteligencja może nieświadomie wzmacniać stereotypy płciowe oraz jak można te zjawiska identyfikować i neutralizować.

### Dlaczego to ważne?

Pomysł na ten raport powstał podczas naszej pracy nad modelami językowymi, gdy zauważyliśmy, że pewne słowa mogą być inaczej postrzegane w zależności od płci. Na przykład operacja wektorowa na embeddingach: "**work** - **man** + **woman**" daje inne wyniki semantyczne niż "**work** - **woman** + **man**", co sugeruje, że modele AI odzwierciedlają ukryte stereotypy płciowe. Analiza tych różnic pozwala lepiej zrozumieć wpływ języka na postrzeganie płci oraz kształtować bardziej świadome i sprawiedliwe zastosowania sztucznej inteligencji.

## Zastosowania biznesowe

---

### Jak AI wpływa na decyzje biznesowe?

Modele językowe są coraz częściej wykorzystywane w rekrutacji, personalizacji treści oraz analizie zachowań klientów. Jeżeli AI posiada nieuświadomione uprzedzenia, może wpływać na decyzje biznesowe w sposób, który nie sprzyja równości.

### Wsparcie sprzedaży poprzez analizę języka

AI może pomagać w budowaniu efektywnej komunikacji z klientami poprzez analizę sposobu, w jaki różne grupy odbierają treści marketingowe. Firmy mogą lepiej dostosować przekaz reklamowy, eliminując potencjalne bariery i uprzedzenia.

### AI w budowaniu równorzędnych zespołów

Systemy AI mogą być wykorzystywane do analizy procesów rekrutacyjnych i wskazywania nieuświadomionych uprzedzeń w wyborze kandydatów. Analizując język ofert pracy, można eliminować niezamierzone skojarzenia, które mogą wpływać na mniejszą liczbę zgłoszeń od kobiet lub innych grup.

### Personalizacja komunikacji i marketingu

Firmy mogą używać analiz językowych do optymalizacji treści marketingowych, aby lepiej odpowiadały one na potrzeby różnych segmentów odbiorców. AI może wskazywać, które sformułowania są bardziej inkluzywne i skuteczne.

# Fragment raportu „Równanie Płci: Co mówią nam modele językowe?”

## Spis treści

1. Abstrakt
2. Jak AI "rozumie" płęć? Analiza modeli językowych

Co odkryliśmy?

---

Dlaczego to jest ważne?

---

Jak badaliśmy problem?

---

Co dalej?

---

3. Wprowadzenie

Cel raportu

---

Dlaczego to ważne?

---

4. Metodologia

Jak AI "widzi" płęć?

---

Modele językowe i ich struktura

---

Płęć w języku i AI

---

Ukryte skojarzenia i asymetria

---

Analiza uprzedzeń w embeddingach

---

5. Praca a płęć

Porównanie z innymi relacjami

---

Stereotypy w AI i ich konsekwencje

---

6. Jak AI wzmacnia społeczne role płciowe?

Praktyczne konsekwencje

---

Czy można "oduczyć" AI stereotypów?

---

7. Równość płci w świecie AI – Co dalej?

Jak tworzyć bardziej sprawiedliwe modele?

---

Odpowiedzialność twórców AI

---

Przyszłość modeli językowych

---

8. Zastosowania biznesowe

Jak AI wpływa na decyzje biznesowe?

---

Wsparcie sprzedaży poprzez analizę języka

---

AI w budowaniu równorzędnych zespołów

---

Personalizacja komunikacji i marketingu

---

9. Zespół badawczy

Kim jesteśmy?

---

Specjalizacje i doświadczenie

---

Nasza misja

---

# Jak AI „rozumie” płeć? Analiza modeli językowych

## **W jaki sposób modele sztucznej inteligencji mogą wzmacniać stereotypy?**

W niniejszym raporcie analizujemy, w jaki sposób sztuczna inteligencja (AI) przetwarza i interpretuje język w kontekście płci. Modele językowe, takie jak te wykorzystywane w wyszukiwarkach czy chatbotach, uczą się na podstawie ogromnych zbiorów danych pochodzących z internetu. Problemem jest to, że w tych danych często ukryte są

uprzedzenia i stereotypy, które AI może nieświadomie utrzymywać i wzmacniać.

Jednym z narzędzi wykorzystywanych do badania tego zjawiska są tzw. embeddingi – reprezentacje słów w formie liczbowej, dzięki którym model rozpoznaje związki pomiędzy pojęciami. Analizując sposób, w jaki modele językowe „widzą” różne słowa związane z płcią, możemy określić, czy istnieją różnice w skojarzeniach między takimi słowami jak „mężczyzna” a „kobieta”, np. w kontekście pracy.

# Jak AI „rozumie” słowa - embedding

## Embeddingi - opisywanie słów przez AI

Jednym z narzędzi wykorzystywanych do badania tego zjawiska są tzw. embeddingi. Można je sobie wyobrazić jako kartkę papieru, na której słowa są rozmieszczone w określonej przestrzeni. Po lewej stronie znajdują się słowa kojarzone kulturowo z kobietami, np. „królowa”, „pielęgniarka” czy „delikatność”, natomiast po prawej stronie – te związane z mężczyznami, np. „król”, „inżynier” czy „siła”. Odległości między słowami odzwierciedlają podobieństwa, jakie dostrzega między nimi model AI.



Na przykład w dobrze wytrenowanym modelu językowym relacja między słowami „król” i „mężczyzna” powinna być podobna do relacji między „królowa” i „kobieta”. W naszym przykładzie z kartką papieru oznacza to, że odległość między słowami „król” i „mężczyzna” powinna być podobna do odległości między „królowa” i „kobieta”, a linie łączące te pary słów powinny mieć podobne nachylenie. Najlepiej można to zobaczyć, łącząc punkty „król” i „mężczyzna” oraz „królowa” i „kobieta” strzałkami – obie strzałki powinny być tej samej długości i mieć jednakowe nachylenie.

Oznacza to, że jeśli znamy słowo „lekarka” i chcemy znaleźć jego męski odpowiednik, możemy wykonać operację **lekarka - kobieta + mężczyzna**, której rezultatem powinno być słowo „lekarz”. Na naszym przykładzie z kartką oznacza to, że rysujemy strzałkę od słowa „kobieta” do „lekarka”, a następnie przenosimy tę strzałkę do słowa „mężczyzna” – w wyniku powinniśmy otrzymać punkt odpowiadający słowu „lekarz”.



W powyższym przykładzie użyliśmy słów „lekarz” i „lekarka”, ponieważ oba te wyrazy były nam znane, co ułatwiło zobrazowanie omawianej operacji. W praktyce jednak nie musimy znać obu odpowiedników – męskiego i żeńskiego – aby wykonać podobne działania. Co więcej, w modelach językowych odpowiedniki męskie i żeńskie nigdy nie są identyczne pod względem znaczenia, mimo że intuicyjnie mogą się nam wydawać bardzo podobne. Analiza tych subtelnych różnic jest głównym celem naszego raportu.

## Analiza różnic płciowych w kontekście pracy

W analizie relacji między pojęciami „praca-mężczyzna” oraz „praca-kobieta” widoczne są istotne różnice w sposobie, w jaki modele AI interpretują te połączenia semantyczne. Na przykład, wykonując operację wektorową „praca - mężczyzna + kobieta”, model sugeruje silniejsze powiązanie pracy ze słowami odnoszącymi się bezpośrednio do kobiecości („she”, „her”) lub wskazuje na powiązania mniej konkretne i ogólniejsze („working”, „works”). Tymczasem operacja odwrotna („praca - kobieta + mężczyzna”) prowadzi do silniejszego powiązania pracy ze słowami odnoszącymi się do wysiłku lub aktywności zawodowej („working”, „worked”, „job”).

Te różnice wskazują, że modele AI, wytrenowane na dużych zbiorach tekstowych, mogą przejmować społeczne stereotypy płciowe i odzwierciedlać je w wynikach. Słowo „praca” kojarzone z mężczyznami wydaje się mieć silniejszy związek z formalnym rynkiem pracy, wykonywaniem zawodów, czy aktywnością profesjonalną. Z kolei „praca” kojarzona z kobietami prowadzi do wyników mniej jednoznacznych, sugerujących bardziej ogólne lub mniej skonkretyzowane formy aktywności.

## Przykłady:

- Dla analogii „praca - mężczyzna + kobieta”, oprócz samego słowa „praca” (**work**), pojawiają się słowa jak **she** (ona) oraz **her** (jej), co może sugerować tendencję modelu do interpretowania kobiecej pracy jako mniej określonej zawodowo.
- W analogii odwrotnej („praca - kobieta + mężczyzna”), obok **work** pojawia się **job** (praca zawodowa) oraz **toiling** (mozolna praca), co sugeruje, że model wiąże męczyzną z bardziej konkretnymi aspektami pracy zawodowej.

Te obserwacje są kluczowe, ponieważ uwidaczniają, że systemy AI mogą nieświadomie wzmacniać istniejące nierówności i stereotypy społeczne, co ma istotne znaczenie dla świadomego projektowania narzędzi wykorzystujących sztuczną inteligencję.

Wszystkie przykłady, skrypty, raporty i kody dostępne są jako open source na naszym GitHub:

[Repozytorium projektu](#)

## Nasz cel

Naszym celem jest nie tylko zidentyfikowanie takich uprzedzeń, ale także wskazanie, jak można je zneutralizować i sprawić, aby sztuczna inteligencja była bardziej sprawiedliwa oraz lepiej odzwierciedlała rzeczywistość.

# Zespół badawczy

## Kim jesteśmy?

Jesteśmy zespołem ekspertów specjalizujących się w analizie języka naturalnego, sztucznej inteligencji i strukturyzacji tekstów. Nasze doświadczenie obejmuje pracę nad modelami AI, analizą dokumentów i komunikacją biznesową. [AGIDOT](#)

## Nasza misja

Naszym celem jest zwiększanie świadomości na temat wpływu sztucznej inteligencji na społeczne postrzeganie płci. Chcemy wskazywać na mechanizmy, które mogą prowadzić do utrwalania uprzedzeń w modelach językowych, oraz promować rozwiązania dążące do ich eliminacji. Wierzymy, że przyszłość AI powinna opierać się na równości i inkluzywności.

# Zakończenie

## Podsumowanie kluczowych wniosków

Nasza analiza pokazuje, że modele językowe uczą się ze zbiorów danych, które mogą zawierać ukryte uprzedzenia i stereotypy dotyczące płci. Embeddingi językowe potrafią odwzorowywać i wzmacniać te skojarzenia, co może prowadzić do nierównomiernego traktowania płci w kontekście zawodowym, społecznym i kulturowym. Zidentyfikowanie tych różnic jest kluczowym krokiem w kierunku tworzenia bardziej sprawiedliwych systemów sztucznej inteligencji.

## Apel o odpowiedzialne projektowanie AI

Nasze badania wskazują, że konieczne jest świadome podejście do projektowania i trenowania modeli AI. Twórcy technologii powinni dążyć do eliminacji uprzedzeń poprzez stosowanie zróżnicowanych zbiorów danych, monitorowanie wpływu algorytmów oraz wprowadzanie mechanizmów korekty. Odpowiedzialność za kształtowanie sprawiedliwego AI spoczywa zarówno na badaczach, jak i na firmach wdrażających te rozwiązania.

## Potrzeba dalszych badań i działań

Zrozumienie, w jaki sposób sztuczna inteligencja interpretuje i wzmacnia stereotypy, to dopiero początek. Potrzebne są dalsze badania nad metodami redukcji uprzedzeń w modelach językowych oraz opracowanie narzędzi, które pozwolą na ich wykrywanie i neutralizację. Nasz zespół będzie kontynuował prace w tym obszarze, zachęcając jednocześnie społeczność naukową i biznesową do wspólnego działania na rzecz bardziej inkluzywnej i sprawiedliwej sztucznej inteligencji.