Algorytmy i miary dla bezstronnych i wyjaśnialnych systemów decyzyjnych

W dzisiejszym silnie skomputeryzowanym świecie, coraz więcej decyzji podejmujemy z pomocą komputera. Pomagają nam w tym *systemy decyzyjne*, czyli algorytmy, które biorąc pod uwagę dostępne dane, podpowiadają najlepszą decyzję w danej sytuacji. Systemy decyzyjne znajdują zastosowanie w takich dziedzinach życia jak marketing (przy podejmowaniu decyzji o tym, jakie oferty pokazać klientowi), w zarządzaniu ryzykiem kredytowym (przy podejmowaniu decyzji o zatwierdzeniu kredytu) czy w ubezpieczeniach (przy podejmowaniu decyzji o tym, czy wniosek powinien być sprawdzony pod kątem potencjalnego oszustwa). W systemach gdzie decyzje są podejmowane w pełni automatycznie, stosowane algorytmy wywodzą się z uczenia maszynowego (poddyscypliny sztucznej inteligencji), gdzie podstawowym zadaniem jest odkrywanie wiedzy z danych i późniejsze wykorzystywanie tej wiedzy w celu podjęcia decyzji. W systemach gdzie ostateczną decyzję zawsze podejmuje człowiek, szerokie zastosowanie znajdują metody z dyscypliny wspomagania decyzji, gdzie celem jest wskazanie użytkownikowi rankingu najlepszych opcji w oparciu o z góry zadane kryteria.

Wraz z rosnącym dostępem do danych, wspomniane wyżej systemy decyzyjne stają się coraz skuteczniejsze, ale też coraz bardziej złożone. W związku z tym pojawia się problem pełnego zrozumienia i kontrolowania systemów decyzyjnych w celu zachowania uczciwości i przyjętych w danym kraju regulacji prawnych. Te problemy leżały u podstaw niedawno zaproponowanego rozporządzenia Unii Europejskiej w sprawie sztucznej inteligencji oraz niedawno przyjętego zalecenia UNESCO w sprawie etyki w sztucznej inteligencji. Problemy z kontrolowaniem systemów decyzyjnych to również szereg wyzwań badawczych dla naukowców, w szczególności związanych z *bezstronnością* i *wyjaśnialnością* algorytmów.

Bezstronność w sztucznej inteligencji oznacza, że algorytm nie jest skonfigurowany tak, aby oceniać decyzje na podstawie płci, rasy, orientacji seksualnej lub innego rodzaju przesłanek wykluczających. Celem jest zapewnienie, że algorytm lub model sztucznej inteligencji jest obiektywny i traktuje wszystkich użytkowników jednakowo, bez dyskryminacji. Istnieje jednak wiele wyzwań związanych z osiągnięciem bezstronności w sztucznej inteligencji, w tym zmiana zachowania algorytmów, gdy występuje niezbalansowanie danych, czyli sytuacja w której posiadamy zdecydowanie mniej przykładów dotyczących jednej z decyzji. Inną trudnością jest definiowanie i mierzenie bezstronności w sposób, który jest spójny i obiektywny.

Wyjaśnialność to zdolność algorytmów do wyjaśnienia swoich decyzji i wyborów w sposób zrozumiały dla ludzi. Wyjaśnialność pomaga w budowaniu zaufania do systemów decyzyjnych i pozwala upewnić się, że są one wykorzystywane w sposób świadomy. Osiągnięcie wyjaśnialności może jednak stanowić wyzwanie, ponieważ wiele algorytmów sztucznej inteligencji, takich jak sieci neuronowe, jest wysoce złożonych i trudnych do zinterpretowania. Ponadto, metody wyjaśniania są najczęściej stosowane do modeli uczenia maszynowego i często są niedostępne dla systemów wspomagania decyzji, które również mogą być złożone i trudne w interpretacji.

Proponowany projekt koncentruje się na problemach omówionych powyżej—wyjaśnianiu systemów decyzyjnych oraz mierzeniu bezstronności. W ramach projektu zbadamy teoretyczne właściwości miar bezstronności i zaprojektujemy nowe miary biorące pod uwagę niezbalansowanie danych. Opracujemy również metody wizualizacji wielokryterialnych systemów decyzyjnych, które pozwolą łatwiej wyjaśniać podejmowane przez nie decyzje. Wyniki naszych badań wykorzystamy w praktycznych zastosowaniach i udostępnimy szerszej publiczności w postaci wyjaśnialnych wielokryterialnych kokpitów decyzyjnych oraz bibliotek dla programistów.