Wśród różnych podejść do sztucznej inteligencji (ang. *artificial intelligence,* AI) szczególnie skuteczne jest uczenie maszynowe. Tym, co wyróżnia niektóre z jego algorytmów, jest problem czarnej skrzynki, który pojawia się przy użyciu nieprzejrzystych poznawczo funkcji. Funkcje te są często zbyt skomplikowane, abyśmy mogli je analizować, na skutek czego pełne zrozumienie kryteriów podejmowania decyzji AI może okazać się niemożliwe. Aby przywrócić zaufanie do zautomatyzowanego podejmowania decyzji, naukowcy koncentrują się obecnie na programie badawczym wyjaśnialnej sztucznej inteligencji (ang. *explainable AI,* XAI), którego celem jest zapewnienie wyjaśnialności, przejrzystości i interpretowalności algorytmów jako sposobów walidacji procesu podejmowania decyzji przez nieprzejrzyste poznawczo systemy inteligentne.

Chociaż powstało na ten temat już wiele prac technicznych, często brakuje im solidnych podstaw koncepcyjnych i nie są one dobrze zintegrowane z badaniami nad *rozumieniem* lub samym pojęciem *wyjaśnienia* w nauce. W tym świetle celem projektu jest filozoficzna analiza problemu czarnej skrzynki w uczeniu maszynowym oraz eksplikacja pojęcia wyjaśnialności AI. Realizacji tego celu towarzyszyć będzie zbadanie relacji między wyjaśnianiem a rozumieniem w nieprzejrzystych modelach AI z wykorzystaniem modeli wyjaśniania rozważanych w filozofii nauki oraz epistemologicznych teorii rozumienia.

Głównym pytaniem postawionym w projekcie jest pytanie o to, w jaki sposób wyjaśnienia pozwalają użytkownikom zrozumieć działanie nieprzejrzystych systemów AI. Ponieważ nie sformułowano dotąd precyzyjnych kryteriów określania, kiedy wyjaśnienia AI są adekwatne, projekt ma również na celu opracowanie zestawu wymagań dla teoretycznie ugruntowanego pojęcia wyjaśnialności AI, które mogłyby odegrać rolę w ocenie modeli XAI. Rozwiązanie tych problemów będzie wymagało znalezienia odpowiedzi na kilka pytań szczegółowych, do których należą następujące zagadnienia: Jakie są kryteria wyjaśnialności AI? Jakie teoretyczne i praktyczne wymagania powinno spełniać „dobre" wyjaśnienie? Jakiego rodzaju rozumienie modeli AI zapewniają lub powinny zapewniać rozwiązania XAI? Na zdobyciu jakiej wiedzy zależy użytkownikom w kontekście interpretacji działania AI?

Do prowadzenia badań empirycznych naukowcy potrzebują pojęć, bez których nie można mówić o wiedzy oraz komunikowaniu jej tym, którzy chcą dowiedzieć się czegoś o rzeczywistości. W związku z tym, podstawową metodą stosowaną w projekcie jest analiza pojęciowa, skupiająca się na identyfikacji i klaryfikacji pojęć w zakresie ich jasności, precyzji, spójności czy empirycznej adekwatności, a także inżynieria pojęciowa, która ma na celu skorygowanie znaczeń istniejących pojęć i wskazanie znaczeń terminów, których powinniśmy używać, aby lepiej naświetlać i rozwiązywać problemy pojęciowo-teoretyczne jakie napotykamy w praktyce naukowej. W ramach badania tej praktyki, istniejące techniczne metody tworzenia wyjaśnień w XAI zostaną poddane krytycznej ocenie oraz zestawione z różnymi rodzajami wyjaśnień rozważanymi w filozofii nauki. Poprzez klaryfikację pojęć, ujawnianie przyjmowanych założeń oraz ocenę praktyk badawczych, filozofia nauki może stanowić ważny wkład do debat naukowych.

W projekcie przeanalizowane zostaną między innymi opracowane w ramach tradycji nowego mechanicyzmu *wyjaśnienia mechanistyczne,* stanowiące opisy interakcji części oraz wewnętrznej organizacji mechanizmów odpowiedzialnych za zjawiska; postulowane przez Daniela Kosticia *wyjaśnienia topologiczne,* opisujące w jaki sposób matematyczne właściwości wzorców połączeń w złożonych sieciach determinują dynamikę systemów wykazujących te wzorce; lub *wyjaśnienia kontrfaktyczne,* mające na celu ujawnianie zmiennych, które powinny być inne w warunkach początkowych zjawiska, aby móc zaobserwować oczekiwany wynik.

Celem wyjaśnienia modelu AI jest uczynienie jego działania zrozumiałym dla użytkowników. Jednak bez wcześniejszego zdefiniowania co oznacza stwierdzenie, że jakaś osoba rozumie dany model lub decyzję, strategiom wyjaśniającym brakować będzie dobrze określonego celu. Dlatego też projekt ma również na celu opracowanie aparatury pojęciowo-teoretycznej dla zagadnienia rozumienia wyjaśnień AI. W ostatnich latach związek między wyjaśnieniem a rozumieniem został szczegółowo zbadany przez Henka de Regta, Kareema Khalifę i Daniela Wilkenfelda, których podejścia można wykorzystać do zintegrowania szeregu strategii wyjaśniających pod względem ich potencjału do wywoływania efektu rozumienia w użytkownikach. Potencjalnie koncyliacyjnym podejściem do problemu rozumienia wyjaśnień AI z uwzględnieniem różnych perspektyw poznawczych użytkowników AI jest także *realizm perspektywiczny* Micheli Massimi.

Wyniki projektu pozwolą sproblematyzować relację między wyjaśnieniem a zrozumieniem w XAI, oferując przy tym nowatorską filozoficzną diagnozę problemu. Oprócz spostrzeżeń teoretycznych, projekt dostarczy także typologii wyjaśnień sztucznej inteligencji oferowanych przez szereg metod zgodnych z literaturą. Wraz z szybkim i powszechnym rozwojem inteligentnych technologii doświadczamy obecnie rosnącego ryzyka etycznego związanego z podejmowaniem nieprzejrzystych decyzji o wysokiej stawce, co może prowadzić do społecznie niedopuszczalnych, a nawet katastrofalnych skutków. Luka w badaniach nad wyjaśnialnością AI powinna być zatem traktowana priorytetowo, aby zapewnić rozwój bezpiecznej i pożytecznej technologii przyszłości. Zadanie to wymaga multidyscyplinarnego podejścia, obejmującego nie tylko informatykę czy nauki ścisłe, ale także filozofię i nauki humanistyczne, w celu opracowania solidnych podstaw teoretycznych dla dalszych prac technologicznych w dziedzinie sztucznej inteligencji.