Normy w językowej komunikacji pomiędzy sztuczną inteligencją a człowiekiem (NIHAI)

Cel projektu: Celem projektu jest identyfikacja w jaki sposób funkcjonuje komunikacja pomiędzy sztuczną inteligencją (AI) i ludźmi, kiedy może stwarzać potencjalne zagrożenia, oraz w jaki sposób można takie zagrożenia zminimalizować.

Komunikacja międzyludzka jest rządzona przez pewne reguły czy normy. Na przykład, normą jest to, że powinno się przekazywać informacje z wiarygodnego źródła. Jeżeli ktoś propaguje fałszywe informacje, to zostanie nazwany kłamcą. Jeżeli ktoś przekazuje niesprawdzone informacje, to straci zaufanie innych i jej słowa zostaną w przyszłości zignorowane jako niewiarygodne. Niniejszy projekt bada czy, gdy ludzie komunikują się ze sztuczną inteligencją, to identyczne, czy też odmienne normy komunikacji znajdują zastosowanie. Ponadto zbadamy, czy te normy mogą być w niebezpieczny sposób wykorzystane do szerzenia dezinformacji, oraz w jaki sposób można przeciwdziałać takiemu potencjalnemu zagrożeniu. Nasz zespół badawczy będzie analizował różne dyskursy, na przykład komunikację pomiędzy dziennikarzami a ich publicznością. Projekt jest realizowany przez badaczy z 4 krajów (Austria, Polska, Rumunia, Szwajcaria) co pozwoli nam na ukazanie także międzykulturowej perspektywy, poprzez gromadzenie danych w wielu krajach.

Opis badań: Projekt badawczy stara się odpowiedzieć za pomocą interdyscyplinarnych narzędzi na 6 pytań. Pierwsze pytanie ma na celu identyfikację ogólnych i szczegółowych norm w komunikacji między ludźmi a sztuczną inteligencją, oraz tego w jaki sposób takie normy różnią się od norm w komunikacji międzyludzkiej. Starając się odpowiedzieć na drugie pytanie, zespół badawczy skoncentruje się na sprawdzeniu, czy manipulacyjne zachowania w komunikacji z AI, takie jak kłamstwa lub zwodnicze zachowania, są postrzegane jako możliwe. Następnie zespół będzie badał, czy ludzie przypisują AI stany mentalne, takie jak na przykład intencje i przekonania, oraz w jaki sposób należy te przypisania rozumieć. Na przykład, jakie są ich konsekwencje dla przypisywania odpowiedzialności. Czwarte pytanie badawcze dotyczyć będzie tego, czy, oraz w jaki sposób, złamanie przez AI norm w komunikacji wpływa na poziom zaufania do AI. Zespół zbada również, czy pewne cechy kontekstualne AI (na przykład AI wyglądające identycznie jak człowiek) mogą wpływać na odpowiedzi na wyżej wymienione pytania. Projekt zakończy się sformułowaniem dobrych zasad odpowiedzialnego projektowania i użytkowania agentów konwersacyjnych opartych na AI. Będziemy odpowiadać na powyższe pytania wykorzystując nowe narzędzia, takie jak na przykład masowe ankiety online, analizę korpusową (bardzo licznego zbioru tekstów), a także wywiady z użytkownikami aplikacji wykorzystujących AI.

Powody, dla których została podjęta tematyka badawcza : Aktualny rozwój technologii sztucznej inteligencji tworzy szereg nowych możliwości. Niemniej jednak, rodzi także nowe ryzyka, takie jak rozpowszechnianie dezinformacji czy manipulacja za pomocą AI używającego języka naturalnego. Projekt ma na celu stawienie czoła temu problemowi poprzez zbadanie, na jakich zasadach opiera się komunikacja między AI a ludźmi, jakie potencjalne zagrożenia rodzi taka komunikacja, oraz jak można te zagrożenia zminimalizować.

Najważniejsze spodziewane efekty : Sformułujemy zasady odpowiedzialnego projektowania i użytkowania agentów konwersacyjnych opartych na AI w celu zmniejszenia potencjalnego ryzyka wynikającego z komunikacji między ludźmi a AI. Będziemy rozpowszechniać wyniki naszych badań wśród szerokiej publiczności oraz publikować artykuły w najlepszych czasopismach naukowych.