Kolokwia psychologiczne PAN

12.06.2026

Dr Artur Gunia

Tytuł:

Zaufanie w epoce generatywnej sztucznej inteligencji: strategie, ryzyka, przyszłość

Abstrakt:

Zaufanie staje się dziś walutą przyszłości sztucznej inteligencji. W epoce generatywnej AI nie jest ono dodatkiem do technologii, lecz kluczowym warunkiem jej skutecznego, odpowiedzialnego i społecznie akceptowalnego wykorzystania. Odwołując się do perspektywy kognitywistyki oraz psychologii zaufania, traktuję zaufanie jako gotowość do podjęcia ryzyka wobec systemu – ludzkiego lub nieludzkiego – którego działania nie kontrolujemy w pełni. Rozumiem je jako proces oparty jednocześnie na przekonaniach (kompetencja i przewidywalność systemu), afekcie (poczucie bezpieczeństwa) oraz normach (oczekiwanie „właściwego" działania). W realiach generatywnej AI zaufanie przestaje być prostą funkcją jakości modelu; staje się rezultatem sposobu, w jaki użytkownicy i organizacje formułują zapytania, weryfikują odpowiedzi, integrują je z własnymi procesami decyzyjnymi oraz biorą odpowiedzialność za ich dalsze wykorzystanie i rozpowszechnianie.

W pierwszej części omawiam zaufanie w relacji człowiek–maszyna, wskazując na fundamentalną różnicę między zaufaniem do klasycznych artefaktów technicznych – gdzie dominują niezawodność i inżynierskie standardy – a zaufaniem do systemów uczących się, które komunikują się językiem naturalnym, generują treści i często funkcjonują jako „czarne skrzynki". W tym kontekście kluczowa staje się kalibracja zaufania: unikanie zarówno overtrust, czyli bezrefleksyjnego polegania na systemie, jak i nieproduktywnej nieufności, która blokuje jego potencjał. Zaufanie do generatywnej AI nie polega na wierze, lecz na świadomym zarządzaniu niepewnością.

W drugiej części analizuję zaufanie w kontekście kreatywności. W projektach twórczych nie chodzi wyłącznie o poprawność odpowiedzi, lecz o poczucie sprawczości, autorstwa, kontroli i sensu procesu. Generatywna AI działa jak współtwórca: inspiruje, przyspiesza, poszerza horyzonty, ale jednocześnie może podważać intuicje dotyczące oryginalności, intencji i „ludzkiej" odpowiedzialności za rezultat. Pokazuję, jak te napięcia redefiniują relację użytkownika z narzędziem oraz jakie strategie pracy – proceduralne, poznawcze i organizacyjne – pozwalają utrzymać zaufanie bez utraty twórczej autonomii.

Trzecia część koncentruje się na czynnikach systemowo obniżających zaufanie: halucynacjach i biasach. Halucynacje przedstawiam jako strukturalne ryzyko modeli generatywnych – zdolnych do tworzenia wypowiedzi płynnych i przekonujących, lecz nieugruntowanych w faktach lub niezgodnych z intencją użytkownika. Biasy analizuję jako ryzyko systematyczne i normatywne: modele mogą reprodukować nierówności i stereotypy obecne w danych, generując treści statystycznie spójne, lecz społecznie szkodliwe. Wskazuję wielopoziomowe

strategie przeciwdziałania tym zjawiskom – od kompetencji użytkownika, przez procedury zespołowe i standardy instytucjonalne, po rozwiązania architektoniczne i algorytmiczne.

W ostatniej części przechodzę do pytania o przyszłość zaufania w świecie systemów coraz bardziej adaptacyjnych. Argumentuję, że kluczowym kierunkiem rozwoju będzie „intent-aware AI" – systemy, które nie tylko generują wyniki, lecz potrafią negocjować znaczenie, zadawać pytania doprecyzowujące, ujawniać poziom własnej niepewności i uzasadniać swoje wybory w sposób zrozumiały dla człowieka. W tej perspektywie zaufanie nie będzie efektem samej skali modeli, lecz rezultatem projektowania całych ekosystemów interakcji: interfejsów, procedur, mechanizmów kontroli i norm odpowiedzialności. To właśnie tam – w architekturze relacji między człowiekiem a systemem – rozstrzygnie się przyszłość wiarygodnej generatywnej sztucznej inteligencji.