Krzysztof Sołoducha
Wojskowa Akademia Techniczna w Warszawie

Projekt spójnej, ekstrapolowanej woli jako narzędzie budowania zaufania do maszyn autonomicznych

Statystyczny paradygmat sztucznej inteligencji odwołujący się do głębokiego uczenia sieci neuronowych spowodował kryzys zaufania do jej wyników, który można poprzeć licznymi przykładami. Problem zaostrza się w sytuacji, kiedy mamy do czynienia z maszynami dążącymi do autonomii, które muszą wypracowywać decyzje na podstawie danych dostarczanych w czasie rzeczywistym, w trybie uczenia bez nadzoru. Problemem staje się nie tylko rozpoznanie sytuacji maszyny w otoczeniu na podstawie wzorców identyfikacji danych, ale także podejmowanie decyzji w oparciu o te rozpoznania. Ich szczególnym przypadkiem są takie rozstrzygnięcia, które zapadają w sytuacji dylematów moralnych – wtedy trzeba do nich zastosować jedną z podstawowych strategii etycznych.

Wobec zasadniczych trudności ze zobiektywizowaniem kryteriów podejmowania decyzji w oparciu o etykę deontologiczną lub też utylitarystyczną, nasuwa się pokusa zastosowania paradygmat statystycznego – odwołującego się do etyki cnót. Próbą wdrożenia w życie tego pomysłu jest koncepcja spójnej, ekstrapolowanej woli Eliezera Yudkowsky'ego.

Przedmiotem mojego wystąpienia będzie zaprezentowanie szczegółów realizacji takiej statystycznej strategii rozstrzygania dylematów moralnych oraz wnioski jakie wypływają z tego projektu dla rozwiązania problemu budowania zaufania do działania maszyn autonomicznych.