Streszczenie popularnonaukowe projektu
Wymóg zrozumiałości systemów uczenia maszynowego wykorzystywanych w stosowaniu prawa

Wyobraź sobie, że zostałeś skazany za przestępstwo. Niedobrze... Trzeba by się dowiedzieć, co przesądziło o takim niekorzystnym wyroku. Gdy tylko sentencja wyroku wybrzmiała, prosisz o jego uzasadnienie. „Nie mogę uzasadnić wyroku, bo nie wiem, na jakiej podstawie system komputerowy go wygenerował" – słyszysz. Brzmi niepokojąco, prawda? Chyba nikt nie chciałby zostać skazany i nie móc się dowiedzieć, jakie były podstawy wyroku. Czujemy, że tam, gdzie jest stosowane prawo, powinno istnieć uzasadnienie podjętych decyzji.

No dobrze – ale co w sytuacji, w której można by zastosować w sądownictwie (albo innej dziedzinie stosowania prawa) pewne systemy: bardzo skuteczne, prawie bezbłędnie określające wszystkie okoliczności spraw. Mogłyby one przejąć część spraw od sędziów lub urzędników, dając im więcej czasu na inne zadania. Systemy miałyby jeden tylko minus: nie dałoby się poznać motywów, na podstawie których podjęły rozstrzygnięcie. Jeśli tak zarysuje się sprawę, nie jest już ona tak oczywista: może warto zrezygnować ze zrozumiałości systemów na rzecz możliwego ogromnego wzrostu efektywności sądownictwa? Ale może nie powinno się iść na żadne kompromisy w kwestii takich wartości, jak transparentność orzeczeń…?

Sprawa jest o tyle trudna, że przedstawione dwa scenariusze tak naprawdę mogłyby być dwoma spojrzeniami na tę samą sytuację. Dlaczego? Otóż najskuteczniejsze i zdolne do wykonywania najbardziej skomplikowanych zadań są często takie automatycznie wnioskujące systemy komputerowe, które mają też najbardziej skomplikowaną strukturę (na przykład głęboką, wielopoziomową – stąd funkcjonująca czasem nazwa „systemów głębokiego uczenia"). Niestety ta skomplikowana struktura sprawia, że przesłanki podejmowania przez nie decyzji są często zupełnie niejasne dla użytkownika. Wie się, że działają, wie się, jaka jest jakość ich działania (a mogą działać nawet niezwykle celnie!), ale nie wie się, co sprawiło, że podjęły konkretną decyzję (nawet jeśli jest się ekspertem) – co jest nazywane „problemem czarnej skrzynki". Ten problem w połączeniu ze znaczną skutecznością takich modeli prowadzi do sprzeczności wartości efektywności (np. ograniczania kosztów) i wartości zrozumiałości sposobu działania systemu (a więc też możliwości uzasadnienia sugerowanych przez niego decyzji). Taka sprzeczność szczególnie uwidacznia się w kontekście prawnym, w którym przecież – jak intuicyjnie czujemy – często pojawia się potrzeba uzyskania uzasadnienia, zrozumienia, wytłumaczenia.

Jak zatem widać, dwa przedstawione na początku scenariusze to po prostu skrajne uwypuklenia największej wady i największej zalety skomplikowanych obliczeniowych systemów komputerowych, które mogłyby być stosowane w sądownictwie (a w niektórych państwach w ograniczonym zakresie już są: np. w USA czy w Chinach). Chcielibyśmy, aby opisana wada objawiała się jak najmniej. Ale do czego – jako obywatele, naukowcy, prawnicy – powinniśmy dążyć? Jaką niejasność modeli jesteśmy jeszcze w stanie zaakceptować, a jakiej już nie? Jaki system będzie zgodny z polskim systemem prawa, a jaki – przez niezrozumiałość działania – nie mógłby być w nim wykorzystany? Jak ważna jest wartość uzasadnialności decyzji prawnych, kiedy możemy ją pominąć, a kiedy powinniśmy uznawać ją za jedną z najistotniejszych? Które modele przewidujące tworzone na podstawie dużych zbiorów danych (czyli modele uczenia maszynowego) dają możliwość uzyskania motywów podejmowanych decyzji, a które nie? I jeszcze: czy jeśli model sam w sobie nie jest zrozumiały dla człowieka, to istnieją techniki, które mogą w kontekście prawnym pozwolić na „wyciągnięcie" z niego przesłanek podjętej decyzji? Na te pytania nie ma prostej odpowiedzi, ale ponieważ wykorzystanie systemów komputerowych w stosowaniu prawa silnie przyspiesza, trzeba pilnie przyspieszyć także badania o tym, jakie standardy powinny spełniać modele uczenia maszynowego wykorzystywane (w przyszłości) w stosowaniu prawa.

Tym właśnie zajmuję się w opisywanym projekcie. Zastanawiam się, jak ważna jest w prawie wartość uzasadnienia decyzji i na ile trzeba ją uwzględnić w tworzeniu systemów komputerowych do stosowania prawa. Badam, które klasy modeli są bardziej, a które mniej zrozumiałe dla człowieka, a także to, jak uczynić te ostatnie w jakiś sposób interpretowalnymi. Rozważam, na ile zrozumiałe powinny być modele wykorzystywane w prawie. To wszystko okraszone tworzeniem w ramach badań przykładowych modeli ucznia maszynowego obejmujących kontekst prawny, czyli powstających na podstawie zbiorów danych dotyczących stosowania prawa (np. o rozstrzygnięciach sądowych). Analiza takich modeli ma pozwolić na lepsze zrozumienie, na ile daje się faktycznie wyjaśnić ich działanie w ramach prawnych przypadków ich potencjalnego wykorzystania (np. przewidywania orzeczeń).