Табела 5.2. Спецификација предмета

Студијски програм: МАС информатика
Назив предмета: Етика и одговорност у вештачкој интелигенцији
Наставник: Сташа Вујичић Станковић
Статус предмета: Изборни
Број ЕСПБ: 8
Услов: Нема предуслова. Пожељно је претходно похађање једног предмета из области вештачке интелигенције или машинског учења.
Циљ предмета: Упознавање студената са етичким, друштвеним и техничким изазовима који прате развој и примену система вештачке интелигенције, са посебним нагласком на три осе које дефинишу савремену расправу: алгоритамску пристрасност (bias), поверење и одговорност у системима вештачке интелигенције (trust and accountability) и јавне страхове и перцепције ризика од вештачке интелигенције (fear and risk perception). Курс комбинује техничке методе (мерење и ублажавање пристрасности, објашњива вештачка интелигенцијеа) са филозофским, правним и друштвеним перспективама, припремајући студенте за рад у окружењу које све више регулише етичке аспекте система вештачке интелигенције.
Исход предмета: По завршетку курса, студент препознаје изворе алгоритамске пристрасности и зна да примени методе за њено мерење и ублажавање, разуме различите формалне дефиниције правичности (fairness) и њихове међусобне компромисе, познаје технике објашњиве вештачке интелигенције (Explainable AI, XAI) и њихову улогу у изградњи поверења, способан је да критички анализира студије случаја у којима су системи вештачке интелигенције изазвали штету, разликује емпиријски засноване и спекулативне ризике од вештачке интелигенције, познаје основе релевантне регулативе (EU AI Act, GDPR) и способан је да у пракси примени принципе одговорног развоја система вештачке интелигенције.
Садржај предмета: - Увод у етику вештачке интелигенције: историјски преглед, етички оквири (последичка, деонтолошка, етика врлина), специфичности вештачке интелигенције у односу на друге технологије, кључни актери (истраживачи, компаније, регулатори, корисници). - Алгоритамска пристрасност (algorithmic bias): извори пристрасности у подацима и моделима, типови пристрасности (историјска, репрезентациона, агрегациона, евалуациона), методе мерења пристрасности и њени стварни ефекти на различите друштвене групе. - Правичност у машинском учењу (fairness in machine learning): формалне дефиниције правичности (демографски паритет, једнаке шансе, калибрација), теореме немогућности и нужни компромиси, технике за ублажавање пристрасности у фазама пре, током и после тренирања модела. - Транспарентност и објашњива вештачка интелигенција (Explainable AI, XAI): интерпретабилност модела, технике објашњења (LIME, SHAP, counterfactual explanations), документациони стандарди (model cards, datasheets for datasets), ревизија и аудит система вештачке интелигенције. - Поверење и одговорност у системима вештачке интелигенције (trust and accountability): шта чини систем вештачке интелигенције достојним поверења, расподела одговорности између програмера, корисника и регулатора, питање одговорности у случају штете, људска контрола над аутоматизованим одлукама

(human-in-the-loop).
- Страхови од вештачке интелигенције и перцепција ризика (fear of AI and risk perception): емпиријске студије о томе како јавност перципира ризике вештачке интелигенције, разлика између непосредних штета (дезинформације, надзор, аутоматизација рада) и спекулативних егзистенцијалних ризика, улога медија и научне фантастике у обликовању перцепције.
- Безбедност и поравнавање система вештачке интелигенције (AI safety and alignment): проблем поравнавања (alignment problem), технике поравнавања (RLHF, constitutional AI), халуцинације и поузданост великих језичких модела, истраживачки приступи у академији и индустрији (Anthropic, OpenAI, DeepMind), критичке перспективе.
- Регулатива и управљање системима вештачке интелигенције (AI governance): EU AI Act, GDPR и заштита података, секторска регулатива (здравство, финансије, правосуђе), процене утицаја (algorithmic impact assessments), стандарди и сертификације, разлике у регулаторним приступима у ЕУ, САД и Кини.
- Студије случаја и шире друштвене теме: историјски примери штета насталих због вештачке интелигенције (COMPAS, Amazon recruiting AI, алгоритми у здравству, препознавање лица), утицај вештачке интелигенције на тржиште рада, приватност и надзор, аутономни системи (возила, наоружање), вештачка интелигенција и креативне индустрије (ауторска права, генеративни модели).

Литература:
1. Dan Hendrycks, Introduction to AI Safety, Ethics, and Society, Taylor & Francis, 2024.
2. Solon Barocas, Moritz Hardt, Arvind Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023.
3. Brian Christian, The Alignment Problem: Machine Learning and Human Values, W. W. Norton & Company, 2020.
4. Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016.
(наставник може изабрати другу одговарајућу актуелну литературу)

Бр. час. акт. наставе: 7	**Теоријска настава: 2**	**Прак. настава: 3**	**Лаб.вежбе: -**	**СИР: 2**

Методе извођења наставе: Фронтални, групни и практични.

Оцена знања (максималан број поена је 100)

Предиспитне обавезе	**поена**	**Завршни испит**	**поена**
активност у току предавања	-	писмени испит	-
практична настава	-	усмени испит	-
колоквијум-и	20	писмено-усмени испит	60
семинар-и	20		