ОТГОВОРИ

на въпроси, препоръки и критични бележки
на Научното жури за конкурса за заемане на академичната длъжност „професор" по професионално направление: 2.3. Философия с тема на конкурса „Етика и регулации" за нуждите на секция „Етически изследвания" на ИФС при БАН,
обявен в „Държавен вестник" бр. 92 от 28 ноември 2022 г.

от доц. д.ф.н. Стоян Андреас Ставру

Уважаеми проф. д.ф.н. Батулева, Председател на Научното жури,
Уважаеми колеги и членове на Научното жури,

Бих искал да изразя своята дълбока признателност и благодарност на всички членове на Научното жури за изразената положителна оценка за приносните моменти, които представих с оглед изследването на взаимодействието между етика и регулации. С настоящето бих искал още веднъж да потвърдя своя последователен и системно изпълняван ангажимент към етическите изследвания и тяхното значение за съществуващите и възможни регулации.

Благодаря за направените бележки и препоръки.

В представените рецензии и становища са формулирани два въпроса в рецензията, предоставена от проф. д.ф.н. Борислав Градинаров и един въпрос в становището на доц. д-р Мариана Тодорова.
Настоящия отговор ще се фокусира именно върху тях.

Отговор на въпроси, поставени от проф. д.ф.н. Борислав Градинаров:

Въпрос № 1:
„Санкционният характер на обезщетението (независимо дали е договорно или деликтно) как кореспондира с безчувствеността на „умната вещ" (в случая самоуправляващия се автомобил)?"

Въпросът на проф. Градинаров е обоснован с посочване на превантивната функция на отговорността, съществуваща „наред с налагането на санкция и поправяне на вредите" и изразяваща се в „недопускане в бъдеще на повторно нарушение от същия характер".
Напълно съм съгласен, че подобна превантивна функция на отговорността, е трудно мислима при интелигентните вещи и софтботите. В най-добрия случай санкцията би могла да бъде възприемана и използвана като своеобразна „обратна връзка", т.е. като

информация, която може да бъде съобразена в процеса на последващо усъвършенстване или самообучение на съответната „умна вещ“. По този начин санкцията не само ще се деморализира: ще съществува напълно независимо от каквото и да е субективно понятие за вина, но ще изгуби и специфичното си юридическо съдържание, свързано с постигането на определени възпитателни и превъзпитателни цели. Тя ще се редуцира до сигнал за необходимост от въвеждането на определена промяна в алгоритъма, въз основа на който вещта взаимодейства със своята среда, включително с обкръжаващите я човешките същества. Превантивната функция на отговорността ще се сведе до по-скоро техническо премодулиране (препрограмиране) на интелигентната вещ по начин, при който поведението й съответства на поставените изисквания. Ефектът от съществуването на санкция в моралните и юридическите правила ще бъде постигнат чрез „техническо решение“, което ограничава или не позволява на съответната вещ да извърши повторно нарушение от същия вид.

Като утвърждаваща се днес концепция може да бъде посочен т.нар. „техноутопизъм“, който наред с множеството си други проявления, има и своите нормативни претенции. Зад тях стои увереността, че правилата могат да се интегрират както във физическата среда, така и в съществуващите в нея интелигентни агенти (т. нар. „нормативна интелигентност“), което да осигури поддържане на предварително зададения (правен/технически) ред. Тази увереност игнорира поне две изключително важни особености, характерни за правните регулации, на които се подчиняват човешките същества:

1) правните регулации не могат да изчерпят предварително всички възможни ситуации и техните „правилни“ решения: това налага преценка във всеки конкретен случай относно това кое е правилното поведение (правилността надхвърля значително като понятие законосъобразността), като тази преценка се извършва при прилагането на по-общи правни принципи и/или чрез използване на различни морални категории; и

2) правните регулации предполагат съществуването на свободни субекти, които могат да спазват, но могат и да нарушават предварително определените правила: изпълнението им зависи от съществуването на субективно генерирана воля относно реализирането на определено индивидуално поведение, която воля се влияе от регулациите, но не и предопределена от тях.

„Умните вещи“ не притежават субективност (воля), наличието на която се предполага от вградените в юридическите наказания мерки за въздействие, промяна и превъзпитание, които имат превантивен характер, т.е. целят предотвратяване на повторно или последващо нарушение на същото правило в сходен контекст. Санкцията при интелигентните вещи по-скоро се изчерпва с „поправяне на вредите“, т.е. определяща е нейната обезпечителна роля, която покрива всички вреди като форма на предварително разпределен риск. В този смисъл може да се твърди, че не става въпрос за санкция в същинския смисъл на думата, а за механизъм за обезвреда, който позволява определен тип обществено търпими и дори полезни отношения да се реализират, въпреки съдържащата се в тях опасност за причиняване на щети. Ако се търси някаква „превантивен“ ефект при този механизъм за обезвреда, той ще е по-скоро под формата на

процедури (протоколи) за „оптимизация“ на съответната интелигентна вещ, т.е. за промяна в нейната юридическа предзададеност чрез съответно редактиране на програмния код, който стои зад нейното взаимодействие с обкръжаващата я среда.

Въпрос № 2:

„Възможно ли е в обозримо бъдеще на изкуствения интелект да бъде предоставено и правото да реши какво би било справедливото обезщетение или наказание, независимо дали отговорността е на хора или на автономни агенти?“

Въпросът е провокиран от тезата, че съвременните технологии могат да се използват за „превръщането на правната норма от текст в таблица“: в „технологично опаковани правила, независими от преценяващия и съдещия субект“. Действително навлизането на генеративния изкуствен интелект (Stable Diffusion, Midjourney, ChatGPT и др.) през последните няколко месеца създава впечатлението, че работата на редица експерти би могла лесно да бъде, ако не изцяло заменена, то поне частично отменена от различни форми на изкуствен интелект. Разработват се различни форми на изкуствен интелект, който на базата на анализ и самообучение върху обобщени и структурирани бази данни от правораздавателни актове предлага проект на съдебно решение по конкретния казус, който в най-голяма степен се вписва в актуалната съдебна практика. Извън сериозния риск подобни „технологични“ решение да възпроизвеждат вече остарели и неадекватни на новата социална обстановка стереотипи, те съдържат в себе си и своеобразна делегация на дейността по правораздаване извън контрола на човека (съдията). Вече не хората „съдят“ за дейността на хората, а често непрозрачни алгоритми, които често са неспособни да мотивират решенията си по разбираем за хората начин.

Макар че в съществуващите към настоящия момент предложения за „правораздаващ“ изкуствен интелект крайното решение е оставено във волята на човека (съдията-човек), убедеността, че обективната безпристрастност е постижима като техническо решение, тласка правораздаването извън сферата на човешкото. Дали обаче вярата, че само „роботите“ могат да съдят при условията на максимално обективност деянията на хората, не е просто една илюзия? До каква степен техните алгоритми са „освободени“ от човешките „изкривявания“ и всъщност искаме ли правосъдие без човешки фактор, т.е. правосъдие без човека като „изкривяване“? Възможността за „операционализиране“ (по подходящия израз на проф. Йотов, използван в неговата рецензия спрямо понятието за справедливост) на справедливостта има своите граници и те трябва да останат в рамките на онова, което е постижимо за хората. Използването на „външна“ инстанция за „окончателно“ решаване на човешки проблеми би могло да доведе до разрушаване на редица морални категории и концепции, които осигуряват легитимността на съществуващите регулации. А превръщането на регулациите (правните норми и етическите правила) в стандарти (протоколи, алгоритми, код) би довело не просто до редуциране на човешкото поведение, но и до загуба на огромна част от неговите субективни измерения.

В този смисъл в отговор на поставения въпрос бих посочил, че дори и изкуственият интелект да служи като възможен ключ за решаването на определен съдебен казус (като своеобразни „вещи лица с изкуствен интелект“ или „изкуствени вещи лица“), той не следва да разрушава правораздаването като човешки акт, т.е. като акт, който е изцяло в и под контрола на хората. Правораздаването е една от областите, в които бъдещите регулации следва да опазят „човешкия суверенитет“ (каквато е и тезата на доц. Тодорова), като всяка „изкуствено интелигентна“ интервенция следва внимателно да бъде обсъждана от хората и да отговаря на съществуващата чувствителност към необходимостта от справедливо и прозрачно мотивиране на съдебните актове. Става въпрос за особено междучовешко взаимодействие „съдещ-съден“, което следва да запази своята специфична йерархичност в рамките на човешкото. Добрият съдия е и си остава човек.

Отговор на въпроси, поставени от доц. д-р Мариана Тодорова:

Въпрос № 3:

„Дали заради пробивите в тесния изкуствен интелект (като ChatGPT на OpenAi) или евентуална реализация на проектите за сингулярност и появата на генерален изкуствен интелект, няма по естествен начин да снемат антропоцентричността и какви биха били последиците тогава за собствеността и екологията, по начина, по който той ги разглежда?“

Въпреки, че на изкуствения интелект се гледа като на (големия, нов) „Друг“, към настоящия момент той често се явява увеличително огледало на редица човешки предразсъдъци, предположения и мисловни навици. Дори и използването на непрозрачни технологии, каквато представлява и т.нар. „дълбоко учене“ (deep learning), когато се отнася до регулирането на определено социално поведение, се базира на статистически данни, генерирани именно от действията на хората. В този смисъл развитието на изкуствения интелект съдържа в себе си не само възможността за „раз-антропо-центриране“, но и за значително усилване на някои „твърде човешки“ модели за възприемане и управление на света. Обсъждането и въвеждането на своевременни регулации би трябвало да има за своя цел балансираното развитие на новите технологии и интегрирането им във вече съществуващите концептуални рамки, даващи представите на човечеството за това, което е добро и позволено, съответно: за това, което е лошо и забранено. Генеративният изкуствен интелект като ChatGPT е „трансформатор“ не само на стрингове (думи, символи и токени), но и на утвърдени форми на комуникация, които в крайна сметка остават центрирани около човека и обслужват негови потребности и интереси. „Мотивите“ зад новите технологии продължават да бъдат генерирани от хората, които ги поръчват, купуват и употребяват.

Тесният изкуствен интелект, независимо от своята възможна „двойна употреба“ (за добро и за лошо), може да бъде мощен инструмент, включително за постигането на екологични цели. Той би могъл да бъде изключително ефективен при събирането, анализа и мониторинга на редица параметри на околната среда, чието поддържане в определени

стойности е ключово за съществуването на човешката цивилизация и на човечеството като цяло. Изкуственият интелект може да генерира различни прогнози за бъдещето, преценявайки определени рискове и пресъздавайки възможни сценарии за развитие на един или друг екологичен проблем. Решенията за настоящето и за бъдещето на човечеството обаче трябва да останат в ръцете на хората, дори и когато това означава „твърде много“ плурализъм, „концептуален хаос“ и неустойчив „многополюсен модел“. Тоталността в решенията на една последна инстанция от най-висш порядък, като каквото може да функционира изкуственият интелект, би обезличила основни морални категории, чрез които човечеството е дефинирало себе си.

Екологичният проблем е преди всичко проблем на хората и решаването му, според мен, няма да „дойде“ от някакво привидно „отвън“ на човечеството, каквато роля би могла да се припише на изкуствения интелект, а в още по-голяма степен и на изкуствения суперинтелект. Струва ми се, че екологичното мислене е форма на „антропоцентризъм с изместен център“: то има за цел да опази целостта на земната екосистема в качеството ѝ на среда, в която човекът може да съществува не само като физическо създание, но и като морален субект. Не смятам, че това е някакъв „лош“ или „скрит“ антропоцентризъм, напротив: това е същностна характеристика на човешкото усилие да запази Земята такава, каквато е, за да продължи той самият да бъде „землянин“ (по думите на Бруно Латур). Признаването на тази човешка цел и балансирането на въздействието, което човекът оказва върху природата, мога да бъдат подпомогнати от развитието на изкуствения интелект, но отговорността е и остава на човека. Това според мен няма да се промени. Поне докато човекът е централната морална инстанция, която създава регулациите.

Още веднъж бих искал да благодаря на всички членове на Научното жури за отделеното внимание върху моя научен труд и за изразените от тях оценки, коментари и бележки.

01.03.2022 Стоян Ставру