# Отговорен изкуствен интелект: етични, правни, социални и икономически аспекти

## Задача №1 "Решаване на етичен казус"

Стефан Димитров Велев, 0МI3400521 спец. "Технологии за големи данни"

### Казус 1. Д-р Уотсън

В тази разработка ще разгледаме проблем, свързван с внедряването на изкуствения интелект в сферата на медицината. Ще очертаем доколко подходящо е да се използват такива технологии в областта и до каква степен може да им се доверим. Всичко това ще дадем под формата на отговор на конкретен етичен казус. Ще разгледаме различните аспекти на контакта между един лекар, д-р Атанасова, и когнитивен компютър, "д-р" Уотсън", през призмата на основните характеристики, които ни определят като личност. За да отговорим на ключовия въпрос дали д-р Атанасова трябва да се довери на "д-р" Уотсън или да последва вътрешния си инстинкт, ще анализираме ситуацията от гледна точна на отделните етични теории – теория на последствията (утилитаризъм), теория на дълга (деонтологична етика), теория на добродетелите (на характера). Ще определим кои страни биха могли да бъдат морални деятели и какви са техните морални задължения в ситуацията на етичния казус. Ще погледнем въпроса от гледна точка на съвременните морални стандарти като прозрачност, отговорност и др. След всички направени анализи ще завършим с гледната точка на автора относно ролята на лекаря в тази ситуация.

Филип не е в добро здравословно състояние. Д-р Атанасова, специалист по неврология в една от найсъвременните болници в София, има съмнения за начални признаци на множествена склероза. От друга страна "д-р" Уотсън, наскоро закупеният от IBM за болнично ползване когнитивен компютър, поставя диагнозата фибромиалгия. Разликата в диагнозите е съществена — докато множествената склероза е потенциално инвалидизиращо състояние, то Филип би могъл да живее с фибромиалгия.

Дали "д-р" Уотсън може да замени д-р Атанасова във формулирането на диагноза поставя по-общия въпрос дали компютърът и в частност системите с изкуствен интелект могат да бъдат морален или дори "хуманизиращ" деятел. Могат ли те да ни научат да бъдем повече хора? Могат ли самостоятелно да поставят диагнози? За да си отговорим на този въпрос, трябва да помислим кое е това, което ни оформя като личности. Една от характеристиките несъмнено е автономността – способността на индивида да определя собствени цели и задачи. В светлината на това качество "д-р" Уотсън би се справил положително, тъй като той е внедрен с цел да "подпомага поставянето на диагнози". В този смисъл неговата цел би била изцяло насочена към това, като той би могъл да вземе самостоятелни решения (задачи) например за допълнителни изследвания, които напълно да помогнат определянето на някаква диагноза.

По отношение на индивидуалността — възможността за персонализиране, имаме както положителни аспекти на "д-р" Уотсън, така и не толкова взимащи предвид отделния случай. Всичко зависи до голяма степен от това как е проектирана системата и с какъв набор от данни е захранена към дадения момент. "Д-р" Уотсън не би могъл да постави точни резултати, ако например в данните няма достатъчно информация за заболяването фибромиалгия. Възможен е и случаят, при който няма достатъчно информация за множествената склероза и когнитивният компютър е търсил друга подобна диагноза. Ако такива данни липсват, то няма как "д-р" Уотсън да подхожда с индивидуалност към отделния случай. От друга страна, ако базата му от знания е достатъчно богата, то той би могъл да поставя допълнителни въпроси, с които да вземе предвид отделни признаци на заболяването за конкретния пациент.

Рационалността е присъща наша характеристика. Способни ли са обаче системите с изкуствен интелект на това? Рационалното мислене свързваме със законите на логиката. Задачата на изкуствения интелект е да генерира логически верни изводи. Подходът на символния изкуствен интелект обаче има големи

недостатъци по отношение на това, че не е възможно всичко да бъде представено толкова лесно под формата на логически твърдения. Необходимо е медицинската информация да бъде много внимателно анализирана, за да може и изводите, които се правят, да отговарят на реалността. От друга страна, рационалното действие е подходът на рационалния агент – да възприема околната среда чрез сензори (sensors) и да действа по отношение на нея чрез задвижващи механизми (actuators). Този елемент е помалко затъпен в етичния казус за "д-р" Уотсън поради спецификата на системата. В случая можем да обобщим, че най-важно е рационалното мислене, което от своя страна зависи от начина на представяне на информацията в базата от знания.

Още една характеристика, която ни определя като личности, е емоционалността — способността на индивида да изпитва емоции. В това отношение "д-р" Уотсън отстъпва на д-р Атанасова, защото той не може да изпитва емоции, освен ако такива правила не са му били предварително заложени. В медицинската сфера е важно лекарят да бъде авторитет, който да вдъхва чувство на сигурност, спокойствие и надежда у пациента, все неща, които една система с изкуствен интелект трудно би могла да изрази истински. По отношение на качествата отзивчивост и изобретателност "д-р" Уотсън би се справил, тъй като възможността за обратна връзка и креативността могат да бъдат програмирани отново най-вече чрез правила и данни.

В крайна сметка установихме, че един от най-големите недостатъци на такива когнитивни компютри с изкуствен интелект като "д-р" Уотсън е качеството емоционалност – нещо, на което ние като хора силно разчитаме в нашето ежедневие. В болнична среда дори е още по-необходимо да имаме уменията да овладяваме емоциите си, а това понякога може да стане само с помощта на лекаря, в контакт с човек. В този аспект може да заключим, че няма как "д-р" Уотсън да се разглежда като морален деятел, а само като инструмент със своите предимства и недостатъци.

При решаването на етични казуси ние винаги вземаме решение въз основа на някаква причина, която надделява за нас над всички останали. Зад изборите, които правим обаче, винаги стои някаква етична теория, въпреки че ние не осъзнаваме това. Ще разгледаме представения етичен казус през призмата на трите подразделения на нормативната етика — теория на дълга, теория на последствията и етика на добродетелите. За целта преди това ще се запознаем накратко с техния фундамент.

Теорията на последствията (утилитаризмът) винаги определя морално правилното поведение чрез анализ на вредите и ползите от последствията от дадено събитие. Ако общите положителни последици надвишават общите негативни последици, то тогава действието е морално правилно. В противен случай, то е морално неподходящо. Според британеца Джереми Бентъм (1748–1832), който е сред първите философи, които систематично формулират утилитарната етична теория, най-важните критерии са социалната полза (social utility) и степента на щастие (happiness), което тя носи. Джон Стюарт Мил (1806–1873) допълнително акцентира към повишаване на нивото на щастие у хората като критерий за обяснение на една морална система. Определя го като универсално, присъщо благо. Според него, ако бъдем попитани защо правим определено действие, ние ще изброяваме причини, които при повтарящи се въпроси "Защо искаш да ...?" ще доведат до отговора "Защото искам да бъда щастлив".

Разглеждаме два типа утилитаризъм – акт-утилитаризъм (act-utilitarianism) и правило-утилитаризъм (rule-utilitarianism). При утилитаризма на действията това, което следим, са последствията от индивидуалните действия (acts), докато при утилитаризма на правилата обръщаме внимание на резултатите от следването на правила (rules) и принципи (principles). По този начин според акт-утилитаризма конкретни действия на изтезание или робство биха били морално допустими, ако социалната полза от тези действия превишава цената на неблагоприятните последици. От друга страна правило-утилитаризмът твърди, че политиките, които позволяват несправедливи действия на изтезание или робство, биха получили като цяло негативни социални последствия от гледна точка на обществото, което само по себе си е неприемливо.

Как да постъпим от гледна точка на нашия етичен казус? Трябва да установим кои последствия ще са по-благоприятни, отколкото неблагоприятни за най-голям брой хора. Ако разгледаме ситуацията от гледна точка на пациента Филип, то диагнозата му, поставена от "д-р" Уотсън, ще бъде по-благоприятна, тъй като той би могъл да живее с фибромиалгия, макар че не съществува конкретно лечение.

Множествената склероза би го поставила в много по-неблагоприятно състояние от гледна точка на последствията, които причинява. Ако се разгледа цялата болница като колектив, то тя активно окуражава лекарите да следват диагнозите, поставени от "д-р" Уотсън, поради опасения, свързани с дела за лекарска грешка, в случай че диагнозата, поставена от лекаря, е грешна. В този смисъл, по-благоприятният изход за ръководството на болницата е да се приеме диагнозата за фибромиалгия на "д-р" Уотсън, но не защото това е по-малкото зло за Филип, а защото по този начин ще се избегнат случаите на възможна лекарска грешка. Трябва да се има предвид обаче, че "д-р" Уотсън има 90% точност при диагностицирането на такива случаи, т.е. все пак има сравнително по-малка вероятност софтуерът да сгреши. В такъв случай, за разлика от противоположния вариант, на "д-р" Уотсън не може да се търси морална отговорност, защото той е машина и не е морален деятел. В такива ситуации отговорността се прехвърля на много хора по веригата, включително и на IBM, откъдето е закупен когнитивният компютър. В известен смисъл това би било по-благоприятна ситуация за болницата, тъй като общественото внимание няма да бъде насочено само към тях, за разлика от случая, в който д-р Атанасова би сгрешила в поставянето на диагнозата си. В крайна сметка според теорията на последствията д-р Атанасова трябва да приеме диагнозата на "д-р" Уотсън, тъй като това води до по-благоприятни последствия за повече хора. У нея обаче остава неприятното чувство, че не се е доверила на вътрешното си усещане.

Теориите, основани на дълга, обосновават морала на специфични, основополагащи принципи на задължението. Тези теории се наричат още деонтологични с оглед на основополагащия характер на нашето задължение. Според немския философ Имануел Кант (1724–1804) то стои в основата на морала, а не последствията от някакви действия, нито пък щастието от това, че сме ги извършили. Нашата рационална природа и капацитет е това, което ни отличава от другите видове същества. Кант разглежда теорията на дълга през призмата на т.нар. "категоричен императив", който в една от версиите си твърди, че трябва да третираме хората като цел, а никога не като средство за постигане на цел, т.е. винаги трябва да се отнасяме към хората с достойнство и в никакъв случай да не ги използваме само като инструменти. По този начин робството няма да бъде никога оправдано не защото може да има негативни социални последствия, а защото по този начин някои хора се използват като средство за постигане на дадена цел.

Теорията на дълга на Кант обаче не ни казва какво да правим, когато две или повече задължения са в конфликт. Една по-нова теория, основана на задълженията, тази на британския философ Уилям Дейвид Рос (1877–1971), ни дава този отговор. Според Рос съществуват т.нар. задължения prima facie (очевидни задължения), които са част от "фундаменталната природа на Вселената" и които ние трябва да следваме. Ако в дадена ситуация само едно от тези задължения може да бъде приложено, ние го прилагаме (то става нашето действително задължение) и знаем как да действаме. Когато обаче възникнат ситуации, в които трябва да избираме между две или повече такива задължения, според Рос ние интуитивно ще разберем кое от тези задължения е нашето действително задължение чрез процес на "рационален интуитивизъм" (rational intuitionism).

Една от теориите, основани на дълга, е теорията за правата. Правата и задълженията са свързани по такъв начин, че правата на едно лице предполагат задълженията на друго лице. Едно от основополагащите за тази теория произведения е "Левиатан" на английския философ Томас Хобс (1588–1679). Чрез това библейско чудовище Хобс описва властта на държавата като суверен над индивида. В началото се намираме в нашето си "състояние на природата" (state of nature), в което липсват съществуващи морални правила и закони. Намираме се в състояние на свобода (естествено състояние), което ни позволява да задоволяваме всички наши желания. Всичко това обаче не е безплатно, защото всеки индивид има различни цели и желания и ние трябва да мислим винаги за това да бъдем внимателни. Липсва чувството за сигурност, намираме се в състояние на "война на всеки срещу всеки". За да излезем от това състояние, идва на помощ "държавата", колективно образувание от хора, в което ние сме готови да жертваме някои от нашите "абсолютни" свободи на суверена в името на това да получим сигурност. Тези идеи се залагат в основата на т.нар. обществен договор, теория, развивана от английския философ Джон Лок (1632–1704) и след това от Жан-Жак Русо (1712–1778), на който не случайно е мисълта "Човек се ражда свободен, а навсякъде е в окови".

След като анализирахме някои от основите на деонтологичната етика, то може да видим как тя би могла да ни помогне и по същество да бъде приложена в нашия етичен казус. Теорията на дълга може да бъде

разглеждана в няколко аспекта. От една страна да приемем диагнозата на "д-р" Уотсън само защото това ще предотврати възможна ситуация на лекарска грешка, което от своя страна няма да доведе до изправяне пред съд, само по себе си представлява нарушение на "категоричния императив". В тази ситуация болницата не би се интересувала по същество от самата диагноза или от това дали има доказателства, че лекарската диагноза е по-точна. За нея е важно във всички случаи да насърчава лекарите да използват системата с изкуствен интелект. По този начин те използват пациентите като средство за сляпо следване на целта си – да избегнат случаи на лекарска грешка. От друга страна д-р Атанасова стои пред две противоречиви задължения – дали да послуша "активното окуражаване" на ръководството на лечебното заведение (защото тя работи там и е длъжна да спазва разпореждания на своя работодател) или да послуша вътрешното си усещане, което и казва точно обратното. Всъщност теорията на Рос ни дава директно този отговор – ние интуитивно ще разберем нашето действително (актуално) задължение, така както в математиката имаме интуиция при доказване на теореми. В такъв случай тя трябва да послуша своя вътрешен глас и да постави диагнозата множествена склероза.

От гледна точка на теорията за правата, внедряването на такива системи все още не е регулирано в България. С публикуването на акта за изкуствен интелект в Европейския съюз се цели да се гарантира, че системите с изкуствен интелект се разработват и се използват отговорно. Правилата налагат задължения на доставчиците и внедрителите на технологии с изкуствен интелект и регулират разрешаването на системи с изкуствен интелект на единния пазар на Европейския съюз. Актът за изкуствения интелект разглежда рисковете, свързани с различните конкретни употреби на изкуствен интелект, като ги категоризира в четири нива на риск и установява съответно различни правила. Системите за диагностициране на заболявания, каквато е в частност и "д-р" Уотсън, се определят като високорискови системи. Такива системи с висок риск ще трябва да отговарят на строги изисквания и задължения преди да получат достъп до пазара на Европейския съюз. Въпреки че регламентът вече е в сила, той все още не се прилага. Поетапното му прилагане започва от 2025 г. Предстои на България да определи кой ще е органът, който ще изпълнява функцията на национален надзорен орган.

Приложимото право, което регулира медицинската дейност в болниците, е Законът за лечебните заведения (ЗЛЗ). Изискванията по отношение на материалната база и апаратурата могат да бъдат намерени в отделни медицински стандарти за качество на оказваната медицинска помощ и осигуряване защита на правата на пациента, които се утвърждават с наредби, издадени от министъра на здравеопазването, на основание чл. 6, ал. 1 от Закона за лечебните заведения. Релевантната наредба за нашия етичен казус е Наредба № 2 от 6 февруари 2014 г. за утвърждаване на медицински стандарт "Нервни болести". В него се определят минималните изисквания към лечебните заведения, които извършват дейности по нервни болести и по детска неврология. Този стандарт, вече на 10 години, естествено, към периода на публикуване не регламентира въпросите свързани с новите медицински технологии и системите с изкуствен интелект. Няма и изисквания по отношение на осъществяването на диагностична дейност от специалистите по неврология. За сравнение в друг медицински стандарт (утвърден с Наредба № 2 от 23 декември 2020 г.) – този по "Обща медицина", може да се намери следният критерий по отношение на изискванията за осъществяване на дейността по обща медицина в първичната извънболнична медицинска помощ:

### Критерий 3.2.13:

Всеки ОПЛ има пълна автономност при вземане на решения за клиничната, промотивната и профилактичната дейност.

#### Индикатори:

- А. Лекарят (лекарите) **самостоятелно** решава към какви специалисти или консултанти да насочва своите пациенти.
- Б. Лекарят (лекарите) **самостоятелно** решава за какви изследвания и диагностични процедури да изпраща пациентите си.
- В. Лекарят (лекарите) самостоятелно определя индивидуалната схема за проследяване състоянието на отделните пациенти.

Като обобщение, все още няма действащи стандарти и изисквания за това как една система като "д-р" Уотсън, която се използва в областта на неврологията, да се използва при определяне на диагноза. Специалистите в рамките на лечебните заведения като д-р Атанасова имат право да вземат самостоятелно своето аргументирано решение освен в случаите, когато политиката на болницата предвижда друго и дотолкова, доколкото тази политика е легитимна (законосъобразна) и може да води със себе си допълнителни задължения.

Последната теория, на която ще се спрем, е теорията на добродетелите (virtue ethics). Много философи вярват, че моралът се състои в спазването на точно определени правила на поведение. Предполага се, че човек трябва да научи тези правила и след това да се увери, че всяко негово действие отговаря на тях. Етиката на добродетелите поставя по-слаб акцент върху последствията (утилитаризъм) и дълга (деонтологична етика) и вместо това подчертава значението на развитието на добри навици на характера. В исторически план теорията за добродетелите е една от най-старите нормативни традиции във философията, намираща своите корени в древногръцката цивилизация, в съчинения на Платон и Аристотел.

Според Аристотел (384 г. пр.н.е. – 322 г. пр.н.е.) не е достатъчно да се научат и възпроизвеждат определени правила, а да се развият и определени добродетели в нашите навици. Правилните добродетели са достатъчни за човек да живее добре. Основният въпрос за това как да се действа в дадена ситуация не бива да насочва към определени правила, както е при други теории, а да се фокусира на това кой би трябвало да бъда. Един човек, на когото са присъщи правилните добродетели, няма дори да си помисли как да постъпи в дадени ситуации, а ще реагира на момента според неговите черти и ценности на характера. Ограниченията, които стоят пред тази теория обаче, са това, че тя е валидна за хомогенни общества, хора с еднакви виждания, мироглед. Такъв е бил случаят с древногръцката цивилизация. В наши дни обаче не може да се каже, че съвременните държави са хомогенни. Напротив, те са хетерогенни, защото хората, които ги населяват, имат различни идеали и гледни точки по различни въпроси, различни ценности, различни добродетели. Да не забравяме една от дефинициите от българския тълковен речник, според която морал е "съвкупност от норми, правила, които регулират, определят поведението на членовете на *дадено* общество". Как разрешаваме конфликти, които могат да настъпят при конкурентни добродетели? Кой от тях надделява? Теорията на добродетелите не ни дава пряко този отговор. Това е и причината с настъпването на "века на разума" и все по-застъпената теза за индивидуалните човешки права, тази теория да остава на по-заден план.

Какво можем да разберем за добродетелите на Филип и д-р Атанасова от описанието на етичния казус? Тук може да се постави въпросът доколко такава система е морално правилно да съществува и с каква база от знания е снабдена. Ако има гаранции за това, че данните са разнообразни и отразяват света по достатъчно приемлив начин, то тогава съществуването на такава система би било оправдано, защото по този начин всеки път, когато лекарите сбъркат с диагнозата, ще има система, която ще даде точен резултат и много пациенти ще бъдат правилно диагностицирани на по-ранен етап, с което ще се даде възможността на повече хора да разгърнат своя потенциал в живота. От друга страна, ако знаем, че все още няма достатъчно обширни данни, които да отразяват до голяма степен медицинските заболявания, то тогава внедряването ѝ и още повече задължаването на лекарите да се придържат към нея би било морално укоримо. И към момента в България текат процеси по дигитализация на системата на здравеопазването, по която има все още какво да се работи. Това е пример за казус, в който няколко добродетеля може да се сблъскат и не може да се вземе решение без единият да надделее над другия. Например, д-р Атанасова може да бъде достатъчно смела и да постави своята диагноза множествена склероза, но може и да бъде достатъчно умерена и да приеме, че 90% точност при диагностицирането на "д-р" Уотсън е значително по-голям процент от нейната точност.

В представеният в разработката етичен казус морални деятели са лекарят д-р Атанасова, ръководството на болницата, както и разработчиците на "д-р" Уотсън – когнитивният компютър от IBM за болнично ползване. Моралните деятели са лица, които действат и вземат решения на базата на морални принципи и етични норми. Те трябва да са морално отговорни за поне част от своето поведение. Имат морални задължения – подлежат на похвала или вина. Трябва да притежават съзнание, способност за рационално мислене, свободна воля, както и елементи на загриженост. Д-р Атанасова има моралното задължение не

само да използва системата с изкуствен интелект, а и да проверява и преценява нейните препоръки въз основа на собствения си опит и знания. Тя трябва да разглежда диагнозите на "д-р" Уотсън не като абсолютна истина, а като допълнителен инструмент, който представлява критична система. Ръководството на болницата също може да се разглежда като морален деятел, тъй като с тяхното решение за внедряване на тази система в конкретната болница, те имат морално задължение пред пациентите, лекарите и обществото като цяло. Те трябва активно да насърчават лекарите от болницата да дават периодично обратна връзка и да докладват случаи на разминавания. Имат тежката отговорност да балансират между иновациите и безопасността, като последното изисква подчертаването на това, че изкуственият интелект е помощен инструмент, а не изцяло заместител на човешката преценка.

От друга страна разработчиците на системата имат моралното задължение да проектират системата, така че тя да предоставя надеждни и обосновани препоръки, които да минимизират риска от погрешна диагноза. Това може да се постигне чрез т.нар. *FAT алгоритми* — мерки за подобряване на справедливостта (*fairness*), отчетността (*accountability*) и прозрачността (*transparency*). Алгоритмичната справедливост идва като решение да се заменят субективните човешки решения с "обективни" алгоритмични методи. Този обективизъм обаче може да остане скрит в наборите от медицински данни, които използваме. За да се разреши този проблем, трябва внимателно да се анализират данните — да се провери разпределението им, да се изчистят, да се обогатят, ако се налага. Ако бъде забелязана някаква тенденция в тях или дори елемент на дискриминация, да се изчакат още нови такива, за да може системата да не накланя везните в една или друга посока. Докато всичко това не бъде разрешено, не бива да се внедрява и да се задължава да се използва една такава животозастрашаваща система в здравеопазването.

Прозрачността е още един важен елемент, който е задължителен в такива рискови системи. Изразява се в това "д-р" Уотсън да може да обясни защо е посочил определена диагноза. Това стои зад идеята на т.нар. eXplainable Artificial Intelligence (XAI) — да се използват алгоритми, които предоставят обяснения за своите предсказания, като кои характеристики или части от данните са повлияли най-много на решението на системата. Друго решение за програмистите е да включват допълнителни модули, които да отчитат доколко сигурен е резултатът, да се дава повече от една възможна диагноза в случай на съмнение и др.

Отчетността е друг ключов компонент за етичната и безопасна употреба на системи с изкуствен интелект. Необходимо е да бъдат въведени механизми за осигуряване на отговорност и отчетност за системите с изкуствен интелект и техните резултати, както преди, така и след тяхното разработване, внедряване и използване. Преди събитието се извършва т.нар. *ex-ante* подход за оценка и анализ, който се фокусира върху очакваните резултати и предвиждания. В случая с "д-р" Уотсън това би включвало множество предварителни анализи по отношение на точността, безопасността, риска, въздействието, направата на етичен и правен анализ, различни предварителни симулации и сценарии. *Ex-post* подходът се прилага след като системата вече е внедрена и работи. При него целта е да се направи пълен одит на системата, който да даде обратна връзка от множество заинтересовани лица – в случая най-вече от лекарите специалисти. За да се подобри нивото на отчетност, целта е *ex-post* подходът да стане част от *ex-ante* процеса, т.е. този одит да може да стане още по време на етапа на проектирането (*by design*). За целта съществуват множество етични кодификации и стандарти, които да ни дават насоки и да задават рамка на нашите действия като разработчици и други специалисти.

В заключение, спазването на *FAT алгоритмите* е от ключово значение за такъв тип системи. Поставяйки се в ролята на разработчици на система, подобна на "д-р" Уотсън, ние трябва задължително да се съобразим с посочените по-горе решения по отношение на трите елемента — справедливост, отчетност, прозрачност. Ролята на лекаря, д-р Атанасова, обаче трябва да бъде незаменим фактор. Специалистите винаги трябва да вземат предвид получените отговори с известна доза съмнение и критичност, за да може да не се стигне до погрешни решения и сляпо следване на инструкции, които са в противоречие с техните вътрешни гласове. Важно е да се отбележи, че системите с изкуствен интелект не са безгрешни и винаги трябва да поставят предупреждение, че има риск от това те да допуснат грешка. Дали можем да си позволим такава грешка — в случая не, защото не трябва да забравяме в каква ситуация се намираме, както и че сме морални агенти, натоварени с цената на обществения договор.