**Тема трета. ЕТИКА НА ИЗКУСТВЕНИЯ ИНТЕЛЕКТ**

***От тази тема ще разберете защо Етиката на изкуствения интелект се разглежда в два аспекта: етични принципи лежащи в основата на използването на решения от ИИ, и етичното поведение на ИИ спрямо хората. Вторият аспект принципно дава ясна диференциация етиката на ИИ от другите цифрови технологии.***

Внедряването на ИИ системите в ежедневието е свързано с много етични проблеми които в следващите години буквално ще ни засипят. Първи пример за такъв проблем е леталния изход от автокатастрофа със самоуправляващ се автомибил Tesla през 2016 г. и Uber през 2018 г., първият протест на разработчици на ИИ в Google отказали да вземат участие в разработването на проект на Военното министерство на САЩ, случаите на манипулиране на достъпа до информация, сексизъм и расизъм в алгоритмите за разпознаване на лица и таргетирането на реклами с използването на ИИ. Големите етични проблеми се появяват при използването на ИИ от държавните служби за контрол над гражданите. Възможните негативни социални последствия при използването на алгоритми в работата на държавата получиха широк отзвук в СМИ държавните органи в някой държави.[[1]](#footnote-1)

Требва да отбележим, че етиката на ИИ силно се отличава от дата-етиката. Спрямо другите технологии сега се обсъждат общи въпроси за промяна на професионалната етика, етиката в използването, етичния отговор на социалните предизвикателства (безработица) и т.н. Специално етиката на ИИ има още много други измерения и те са свързани с поведението на системите в ситуациите касаещи хората.

***Принципно важно е, че системите на ИИ могат:***

* ***Самостоятелно да вземат решения касаещи човека;***
* ***Да анализират данни в такъв обем и скорост с които човек не може - следователно той не може да провери тяхната достоверност!***

***От тук проблемът е да се определи до колко решенията взети от Интелектуални автономни системи (ИАС) не влизат в разрив с етичните норми.***

От тук можем да говорим за два корено различни аспекта на ИИ.

Първият аспект предполага, че първичните програми за ИИ са написани от човек, но в развитие системите на ИИ започват да имат самостоятелност и сътветно като всеки разум ще искат да се самоусъвършенствуват, преструктурират и да увеличат своите параметри. Създавайки системи вземащи критично важни решения за човека, съответно ние като хора искаме да получим гаранции, че те се ръководят от етични съображения и това изисква привнасянето им при създаването на ИИ. Усложненията идват от това, че моралният избор не се прави на основа на ясно маркирани норми и закони, а се формира от правила, принципи и субективни мнения базирани на оценката кое е „добро“ и кое е „лошо“. Такава база е изключително трудно да бъде заложена в ИИ.

Вторият аспект на етиката на ИИ предполага анализ и предотвратяване на етични колизии, възникващи в процеса на използването на ИИ: нарушение на приватността, дискриминация, социално разслоение, проблеми с трудовата заетост и т.н. Съвсем регулярно се поставя и въпроса за професионалната етика на разработчиците на системи за ИИ. Несъмнено отказът от разработването на системи на ИИ е невъзможно и надпреварата в тяхното създаване вече върви с огромни темпове.

**Формулировка за Изкуствен интелект?**

**Проблем в сферата на етиката на ИИ е липсата на единна понятийна база. До колкото няма формално или конструктивно описание на основни етични постулати в областта на техническите изследвания, моралните аспекти се свеждат до битовото, интуитивно разбиране. Самите специалисти в областта на философията не винаги си представят достатъчно ясно дадено постижение в областта на технологиите и ИИ. Очевиден е разривът между разработчиците, изследователите и философите.**

Понятието „Изкуствен интелект“ семантично предполага комплекс от технологии и наука и в този учебник ще бъдат използвани и двете значения.

*„Комплекса от технологични решения позволяващи имитацията на когнитивните функции на човека включващи самообучение, вземане на решения без предварителен алгоритъм, получаващи като резултат конкретни резултати съпоставими с резултатите от интелектуалната дейност на човека приемаме като формулировка за Изкуствен интелект“*

Комплекса от технологични решения включващи в себе си информационно комуникативна инфраструктура, програмно обезпечение включващо машинно обучение, обработка на данни и търсене на решения реализирани с помощта на съответните процесори и сървари.

Изкуствения интелект е наука и технология включваща набор от средства позволяващо на компютъра на база натрупани знания да генерира отговори на въпроси и да прави експертни изводи без те да са заложени в него предварително от разработчика. Областта дефинирана като „Изкуствен интелект“ е междудисциплинарна и влиза в комплекса на компютърните науки и създадените от нея технологии определяни към информационните технологии.

*Като рождена дата на ИИ като научна област се приема 1956 г. в САЩ където на двумесечен семинар в Дартмут организиран от Джон Макарти, Марвин Мински, Клод Шенон и Натаниел Рочестър са поканени учени от еисок ранг изучаващи теориите за управление, автоматика, невронни мрежи, теории на игрите и изследванията на интелекта (Артур Самуе, Алена Нюел, Херберт Сеймон, Тренчард Мур, Оливер Селфридж идр.)[[2]](#footnote-2)*

Прието е да се диференцират две групи проблеми подлежащи на изследване на ИИ като наука:

* ***Силен (интегрален, универсален) хипотетичен ИИ способен не само да решава интелектуални задачи,а и самостоятелно да си дефинира цели сравними с човешкия интелект или превъзхождащи го. За проблемите на Силния ИИ е необходимо да има високо ниво на познание относно човешкия мозък. Невробиологията е натрупала огромно количество емпирични знания за анатомията и физиологията на мозъка и неговите молекулярни и генетични механизми. Общия принцип на преработка на информация в мозъка не е до край изяснена – единствено категорично ясно е, че се случва по съвсем различен начин от този на компютъра. От там и създаването на Силен ИИ е твърде неопределено. Още през 1975 г. някой учени прогнозират създаването ИИ сравним с човешкия около 2000 година. В последствие датата на евентуален научен пробив е отсрочена за 2045 г.***
* ***Слаб (приложен) ИИ изграден с методите и програмните системи решаващи отделни интелектуални задачи. Тук успехите са много по-значителни*.**

Определението за интелектуална задача я дефинира като такава за която човекът няма алгоритъм. Ако човек изпълнява дадена задача по алгоритъм винаги получават един и същ резултат. Когато се решава интелектуална задача хората използват своите знания, умението да разсъждават и съобразителността си (всичките са различни при различните индивиди). Главният успех на ИИ през 60-те години се състои във формализацията на тези интелектуални възможности - разработването на методи за извеждането на знания , моделиране на съжденията, еврестично търсене и т.н.

Паралелно с това се развиват и други направления на ИИ на принципите на машинното учене и технологията на невроните мрежи. С разширяването на областите на използване на тези технологии нарастват съответно и очакванията към тях въпреки, че невротехнологиите имат редица ограничения – трудности при оценяването на качеството на обучаващата се извадка водещо до трудности в тълкуването на получените резултати.

Трудности възникват при създаването на безпилотни системи (особено автомобилни) и това не допуска масовото им производство. При тях особено проблемна е оценката на ситуацията (отделянето на фигурата от фона, движещи се от неподвижни предмети, източниците на опасност). Решаването на тези проблеми може да помогне на машинното обучение на данните и процедурите. Проблемът се състои в това, че има ситуации при които обучаваща извадка изобщо не се прави. видовете интелектуални системи (ИС) биват:

|  |  |
| --- | --- |
| **Видове ИС** | **Описание и използване на ИС** |
| Управляващи интелектуални системи | Управление на производства, проектиране, бордови интелектуални системи в авиацията |
| Динамични роботехнически системи – роботи и безпилотници | Робот – автономна движеща се система с дистанционно управление или с вградена програма. Съществуват роботи с адаптивно поведение, способни да преодоляват препятствия, да „дават оценка за ситуация и да се ориентират в местност. |
| Полиагентни системи | „колективни“ роботи“ с обща цел (изследващи дадена територия). Управлението им е „външно“ и „вътрешно“ с помощта на робот координатор който може да е част от „колектива“. Има вариянт при който координатор не е необходим, а се осъществява ръководство по хоризонтала (между всички). |
| Системи за поддръжка при вземане на решения | Едно от направленията са т. нар. „когнитивни карти“. Такива схеми описват някакви ситуации и връзките между тях, а също и до колко един фактор влияе на друг. С тяхна помощ може да се направи разчет примерно как повишаването на данъците ще повлияе на бюджета и цените. Такива инструменти позволяват проследяването на последствията от дадено решение , но е важно да се отбележи, че самата система не взема решения,а само дава препоръки – отговорността е изцяло на човека. Такива системи масово се разработват в различни области |
| Когнитивни изследвания и когнитивни модели | Опити да се формализират познавателните процеси на човека. Такива изследвания имат своето продължение във „Силните системи ИИ |
| Инженерни знания и онтология | Експертни системи ориентирани към тиражиране опита на специалистите в области където качествените решения зависят от нивото на експертизата и от важност е опита на специалиста (медици, юристи, икономисти). Много компании създават Експертни системи за вътрешно ползване в ключови за тях технологии |
| Моделиране на съждения | Тук се използва логиката, но да бъде тя формализирана е трудна задача и най-вече търсенето на здрав смисъл. В разсъжденията си човек използва разпознаването, опита, спомена за нужния прецедент. Моделирането на съжденията включва: моделиране на съжденията на основа прецедентна аргументация или ограниченията, моделиране на съждения свързани с неопределеност, генериране и проверка на хипотези и др. |
| Обработка на естествения език | Машинен превод на текстове, анализ на текстове с търсене на нужната информация, класификация по съдържание, автоматично рефериране, разпознаване, превод и генериране на изказ. Повечето от тези задачи стават изпълними с помощта на машинното обучение. |

Освен това елементи на ИИ се използват в съвременни цифрови технологии като киберфизическите системи („Умна къща“), допълване на реалността и пр.

**Етично използване на изкуствения интелект**

**Внедряването на ИИ и други цифрови технологии е възпрепятствано от ниското ниво на гражданско доверие към алгоритмите и новите технологии по принцип, а в същевременно липсват разбираеми етически рамки в прилагането на ИИ. В насточщта тема ще бъдат дадени общите характеристики и перспективните подходи за решаване на етическите казуси свързани с използването на ИИ:**

* **Отговорност за етично или неетично поведение на ИИ при вземането на погрешни решения, нанасяне на щета поради грешки и т.н.**
* **Предубеденост на алгоритмите (bias);**
* **Обезпечаване и регулиране на ИИ (обяснителна компонента);**
* **Проблема за приватността при използването на ИИ;**
* **Надеждност на технологиите в ИИ;**

**Отговорност на Изкуствения Интелект**

Проблема за отговорността за действията на изкуствения интелект за първи път е поставен при прилагането му в автомобили без водач и завършва с хипотетичните системи за вземане на решения от планетарен мащаб – тема която е дай дискутирана в момента. Проблемите са формулирани в областта на доверието, където на хората им се налага да залагат на системите ИИ като се има в предвид автомобилният транспорт, фармацефтиката, медицината, образованието и т. н. Институцията на отговорността при различните типове роботи има различни нюанси в зависимост от обществената сигурност, контрулируемост и способността за обучение. Друг аспект в много от случаите е много трудно да се установят обстоятелствата при нанасянето на вреда и в една и съща ситуация се получават различни решения от гледна точка на конкретна юристпроденция. При обсъждането на различните принципни подходи при установяването на отговорност за действията на ИИ вариантите са:

* Пълно освобождаване на ИИ от каквато и да е отговорност за действията им;(аналогично с действията на непреодолима сила)
* Частично освобождаване от отговорност (освобождаване на конкретни лица от каквато и да е отговорност и същевременно изплащане на пострадалите на компенсация за понесени вреди от различни източници);
* Отговорност за вина настъпила само в зависимост от вината на конкретния субект (прозводител, разработчик, лица отговарящи за обучение на ИИ, собственик, ползвател и пр.)
* Отговорност без вина на определени лица (производители) по общите правила възприемани като отговорни за действията на ИИ.
* Лична отговорност на роботите при условия, че са натоварени с правосубектност - права и задължения при наличие на статус на електронна личност

Поради факта на дискусионност и липса на общо решение на учени и разработчици можем да представим различните гледни точки:

1. В етиката има понятие „споделена отговорност“ – отговорността е разделена между разработчика и собственика или ползвател и самата система ИИ, ако e зададена възможности за допълнителното `и настройване.
2. В повечето случаи с отговорността е натоварен „програмиста“ въпреки, че той е само оператор залагаща правилата на поведение делегирани от експерти и специалисти в дадена област (в етиката съответно философ). Морални правилно е, отговорността да се носи от експерта предоставящ етичните правила за действието на системата инплантирани в ИИ.
3. Третия вариант е търсенето на отговорността освен от философа и разработчика и от законодателя, т. к. отговорността се делегира в правов режим.
4. Друга гледна точка говори за това, че както и при всяка техническа катастрофа се търси кой е виновен (примерно при авиокатастрофа, се назначават експерти да решат чия е вината, на производителя, оператора, авиокомпанията или пилота). Затова от голямо значение е валидизирането на система за отговорност по отношение на ИИ.

Ако трябва да се обобщи, то трябва да кажем, че самата концепция за изкуствен интелект е нееднозначна и от там на неопределеност е подложена и формулировката за отговорността. Като универсално се оформя само правилото: *Отговорен е човекът!* Тук трябва да се избегнат двете крайности в случаите: когато отговорността не се носи от никого и тогава когато цялата отговорност се стоварва върху изкуствения интелект.

**Степен на прозрачност на изкуствения интелект**

Системите ИИ са способни да се самообучават, самоусъвършенстват и развиват. Въпросът как една система на ИИ взема решение е от особенна важност, т. к. всичко това става като резултат от сложен алгоритъм. Действията на самия ИИ трябва да имат видимост за широк кръг от заинтересовани страни поради следните причини:

* Тя е важна за ползвателя до колкото формира доверие към системите, предоставяйки достъпно разбираем начин за това какво и как се извършва в нея.
* Сертифицирането и валидизацията на подобна прозрачност на Интелектуалните Автоматизирани Системи (ИАС), за да може да се прави проверка за съответствието на процесите изпълнявани от тях с действащото законодателство.
* ИАС трябва да са прозрачни при разследване на аварии, за да може да бъдат проследени вътрешните процеси довели до нея. Такава достъпност е необходима и на адвокати и експерти.
* Революционните технологии на безпилотните автомобили трябва да бъдат достъпни и прозрачни за широката публика, за да повиши доверието на обществото към тази технология.

Примерно: Канадската емиграционна служба е разработила през 2014 система за автоматизация на дейността си. Алгоритмичните системи се прилагат на всички етапи при емиграция в Канада при това правителството по никакъв начин не дава достъп до това какви критерии се използват при оценката на емигрантите и бежанците, какъв тип данни се въвеждат в автоматизираната система, кой има достъп до информацията и как тя ще циркулира в другите ведомства, до колко правителството приема като допустими грешките в системата[[3]](#footnote-3).

Алгоритмите на работа на невронните мрежи е изключително сложна за интерпретация и следователно резултата от тяхната работа могат да бъдат подложени на съмнение и отхвърлени от човека. Отсъствието на разбиране за това как изкуствения интелект постига своите резултати е една от причините за ниското доверие към съвременните технологии и това рефлектира върху неговото развитие.

Трябва на всяка цена да отбележим, че не може да има и пълна прозрачност за изкуствения интелект – степента на прозрачност за едни и същи системи зависи от тяхната функционалност, целева аудитория и т. н.

Както всяка интелектуална собственост, уникалността на всеки алгоритъм създаден от даден разработчик трябва да бъде запазен в тайна в дадена степен (освен договорените от разработчика). Обяснителната компонента трябва да съществува в такъв вид в който да показва хода на работа на системата без да показва всички „механизми“ на собственото си функциониране. Експертните системи трябва да показват всички вериги на съжденията си на ползвателя, а анализаторите на интелектуални данни трябва да представят своите хипотези в разбираем за човека вид. При ИИ на основата на невронните мрежи, до този момент обяснителната компонента не може да бъде реализирана, но не е далече моментът когато и това ще бъде възможно.

**Предубеденост спрямо алгоритмите**

Това е един от най критичните проблеми пред системите за ИИ – незабележими на пръв поглед диспозиции и предположения могат да останат скрити в базата данни при въвеждането за обработка и на тяхна основа системата наслова някакви резултати които влияят на нейната обективност и правят взетите решения предубедени. Такива „слепи петна“ се реплекират непрекъснато на всички нива на обработката и резултатите и имат комулативен ефект. Справедливостта на алгоритмите е един от основните проблеми в създаването на етичен ИИ.

Самият ИИ обаче може да помогне при преодоляването на предубедените решения[[4]](#footnote-4). Според изследванията алгоритмите могат да намалят расовото и етническо неравенство в системите на правосъдието[[5]](#footnote-5).

Аналогично автоматизираните системи на финансов андерейтинг могат да бъдат полезни за кандидати с недооценен кредитна история.Положителна черта на ИИ отличаваща го от човека е, че решенията които взема са достъпни за анализ и детайлна проверка. Етичните нарушения се получават при некоректно събиране на данни , непълна анонимизация или недостатъчна валидизация на входните данни използвани при обучението на системите на ИИ. Неволна причина може да бъде и селектираната информация водеща до дискриминация.

*Изследвайки кредитната история при вземането на решение за наемане на работа е възможно да се нанесе вреда на социално незащитени граждани, въпреки наличието на връзки между качеството на кредитната история и поведението като работник няма доказана релацция[[6]](#footnote-6). В САЩ има задействана програма за прогнозиране на престъпления PredPolе обучена с етнически изкривена извадка, при което полицията бива изпращана много по-често към адреси на етнически малцинства[[7]](#footnote-7). Обучената на полуистински истории на заболявания програма IBM Watson понякога дава смъртоносни препоръки при лечението на рак.[[8]](#footnote-8)*

Създаването на справедлив ИИ е продукт на дълги дискусии, изследвания и стандартизация. На първо място е търсенето на откриването и намаляването на човешките предубеждения.

*Специалистите от лондонската компания Deep Mind предлага като защита от въздействието на човешките предубеждения да се използва методът „хипотетична справедливост“(counterfactual fairness)[[9]](#footnote-9). За да се формира справедливо и непреднамерено съждение за гражданите ИИ формира хипотетична ситуация в която въпросните граждани притежават противоположни признаци – жените се представят като мъже, бедните като богати, афроамериканците в бели и т.н. По този начин реалния статус не влияе върху дейността на гражданите и е освободен от предубеждения и съответно справедливо.*

Другата възможност е усъвършенстването на самите системи ИИ започвайки с начините за използването на данни и завършвайки процесите разработването им. От значение е и внедряването и използването, за да може да се укрепи индивидуалните и обществени нагласи или предотвратяването на предубежденията и свързаните с тях проблеми. Междудисциплинарното сътрудничество е призвано да осигури разработки, технически иновации и методи на работа и етически стандарти. Практиката показва, че това е възможно.

*Amazon спира използването на системи за подбор на персонал след като намира грешки в алгоритъма свързани с джендърни предрасъдъжи. Алгоритъма разпознава шаблонни думи в резюмета, а не професионални навици. Изходните данни служещи за обучение на системата се оказват резюмета основно на бели мъже. Алгоритъмът изключва резюмета съдържащи думи използвани от жени. В резултат се появяват диспозиции по отношение на жените при кандидатстването им за работа.[[10]](#footnote-10)*

Диспозициите е възможно да бъдат не само естествени акумулирани от специфика на данните, а и изкуствено създадени съобразно интересите на заинтересовани лица. Такава диспозиция може да определи дребни изменения в маршрута на ползвателя примерно в навигационна система за да може той да „посети“, а не да подмине дадена дестинация (магазин, билборд и пр.). алгоритъмът може да предизвика и статистическа корекция която е социално неприемлива или незаконна –моделите на ипотечните кредити откриват, че възрастните хора по-често не спазват графиците на вноските и на тази основа ги изтласква в класацията за получаване на кредити.

*В САЩ алгоритмична одитна програма за прогнозиране на обема на медицинската помощ открива диспозиции по отношение на афроамериканци. Независимо от единна методика за всички пациенти, алгоритъма приема цветнокожите пациенти като по-малко нуждаещи се от медицински грижи от колкото белите, въпреки очевидната потребност на първите. В самия код не е заложен приоритет за белокожите пациенти и алгоритъмът работи нормално. Грешката е на разработчика залагайки равни разходи за медицинска помощ поради еднаквата потребност от нея, но на база разходи в миналото без да отчита факта, че тази потребност зависи от доходите и социалния статус. Така се закрепва дискриминация заложена далече преди създаването на алгоритъма, т. к. в миналото потърпевшите са ползвали по-малко медицински услуги поради ниските си доходи[[11]](#footnote-11).*

Какъв е начинът да се преценява справедливостта, за да се минимизират предразсъдъците?

Проблема идва от това, че самото понятие справедливост изисква конкретно дефиниране и изследване поради факта, че има поне 20 определения за нея, но нито едно не е достатъчно изчерпателно. Обикновено се фиксира индивидуалната справедливост без да се взема под внимание на груповата такава – необходима е различна система от показатели и стандарти в зависимост от обстоятелствата и варианти на използване.

Важна съставляваща на справедливостта на системите ИИ е прякото участие на човека. Въпреки стратегическите показатели на справедливостта, безусловно са адекватни, но не отчитат нюансите на социалните условия в които се развиват системите на ИИ и свързаните с това събиране на данни.

Разбира се това поражда редица въпроси:

* Къде и под каква форма са необходими човешките съждения при разработването и експлоатацията на ИИ?
* Кой решава кога системите на ИИ вече са минимизирали диспозициите и са готови за безопасна експлоатация?
* В кой ситуации е напълно възможно автоматизирано вземане на решения?

В момента нито един алгоритъм на оптимизация не може да взема решения самостоятелно и не трябва да се поверява на машините – необходими са човешки съждения базирани на междудисциплинарен анализ свързващ познанията на хуманитарните и социалните науки, правото и етиката. Института McKinsey препоръчва високо диференцирани препоръки в работата с ИИ за намаляване на диспозициите адресирани към специалистите и ръководителите на системи за ИИ.

Диспозициите при вземането на решения от хора и машини не само водят до пагубни последствия за хората подлагайки ги на дискриминация, но и носят загуби загражданите и държавата, като необосновано ограничават възможностите на отделни субекти да участват и да имат своя принос в икономиката и обществото. Минимизирането на диспозициите в системите на ИИ е критична по отношение на доверието към тях и допринася те да могат да реализират своя потенциал за държавата, икономиката повишавайки ръста на производителността и решавайки редица социални проблеми.

**Надеждност на изкуствения интелект**

Проблема за безопасността и надеждността на ИИ изглеждат откъснати и далечни по отношение на етиката, но въпреки това имат непосредствено отношение към нея трябва да бъдат разгледани в няколко аспекта.

Първият е определено технически засягайки надеждността и безопасността на програмно техническите системи въобще. В обикновените технически системи има по-голям риск от грешки, и повреди за разлика от интелектуалните и затова този аспект не е специфичен за ИИ (грешките в системите за автоматично управление на атомните електроцентрали са много по-опасни от колкото тези на ИИ при вземане на решение за потребителски кредит).

Вторият аспект засяга самата същност на работата на интелектуалните системи. До колкото ИИ вземат решения на слабо формални задачи използвайки евристика (трудно доказуема коректност на правилата практически даваща приемливи резултати), правдоподобни разсъждения и механизми. Това сваля очакванията от системите на ИИ да взема оптимални и еднозначно верни решения и създава възможност за намиране на решения които са „субоптимални“, „разумни“ или „пригодни“ за съответния случай. Тук има пресечна точка между надеждността и прозрачността на ИИ – разработчика може бързо да намери грешката. В непрозрачните системи наричана „черна кутия“ има възможност за внасянето на скрити корекции делегиращи вземането на решения дълго време по различен начин без да предизвиква съмнение.

Третия аспект е наречена проблемна програмите съветници („опит оператори“) имащи пряко отношение към етиката. Проблемите са актуални не само за интелектуалните системи защото с тяхното прилагане те се издигат на съвсем друго ниво. Експерименталните изследвания показват, че в условия на неопределеност и дефицит на време в човека се създава свръх доверие към ИИ и роботите, т. е. те вярват повече на машините от колкото на себе си[[12]](#footnote-12). Доверието още повече нараства ако програмата дава коментар на своите действия. Изцяло разчитайки на системите на ИИ човек все по-рядко взема самостоятелни осъзнати решения което крие риск от грешки и загуба на квалификация. притежателя на интелектуална система е изкушен да не наема ниско квалифицирани специалисти или да предлага по-ниско заплащане. От тази гледна точка е важно в опасните производства, медицината и образованието специалистите да запазят високата си квалифицираност – не е важно те да знаят точно кое копче трябва да натиснат, а да познават системата с която работят детайлно. Колкото и да е „умна“ интелектуалната система, на второ ниво е необходим човешки контрол и хората трябва да са достатъчно подготвени, за да го осъществяват.

Стандартизацията трябва да е способна да повишава нивото на надеждност на системите отчитайки минималното ниво след което системата губи своята надеждност. Това естествено не трябва да се првръща в догма която спира иновациите в технологията.

**Приватност на информацията получена с помощта на анализ на метаданни**

Алгоритмите на ИИ имащи възможността да получават нова персонална информация за хората се случва през анализ на голяма база данни извлечени от метаданни. Събирайки повече данни за човека в цифров профил собственика на алгоритъма (фирма, държавна служба, полиция) получава възможност във все по-голяма степен да предскаже поведението на субекта(какво ще потребява, за кого ще гласува, как може да бъде манипулиран и пр.).

В наше време държавата и обществото волно или неволно трябва да формира своето отношение към допустимия праг на обработка на такива данни на базата приватност, прозрачност и справедливост на ИИ създавайки стандарти за разработване на етични интелектуални системи.

1. Simonite,T. AI experts want to end «black box» algorithms in government // Wired Business.

   URL: https://www.wired.com/story/ai-experts-want-to-end-black-box-algorithms-in-government [↑](#footnote-ref-1)
2. McCarthy J., Minsky M.L.,Rochester N. Et al. A Proposal for the Dartmouth Summer Research Projecton Artificial Intelligence// Stanford.URL:http://www formal.stanford.edu/jmc/history/dartmouth/dartmouth.html [↑](#footnote-ref-2)
3. Kenyon M. Bots at the Gate: a human rights analysis of automated decision-making in Canada’s immigration and refugee system // The Citizen Lab. URL: https://citizenlab.ca/2018/09/bots-at-the-gate-human-rights-analysis-automated-decision-making-in-canadas-immigration-refugee-system/ [↑](#footnote-ref-3)
4. Tackling bias in artificial intelligence (and in humans) // McKinsey. URL:https://www.mckinsey.com/ featured-insights/arti너cial-intelligence/tackling-bias-in-arti너cial-intelligence-and-in-humans [↑](#footnote-ref-4)
5. Kleinberg J.,Lakkaraju H., Leskovec J.et al. Human decisions and machine predictions// The Quarterly Journal of Economics. 2018. Vol. 133, no 1. P. 237–293. [↑](#footnote-ref-5)
6. Silberg J., Manyika J. Tackling bias in artificial intelligence (and in humans) // McKinsey Global Institute. URL: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans [↑](#footnote-ref-6)
7. Lum K., Isaacn W. To predict and serve?// Significance. 2016. Vol. 13, no 5. P. 14–19. URL: https://doi.org/10.1111/ j.1740-9713.2016.00960.x [↑](#footnote-ref-7)
8. Ross C., Swetlitz I. IBM’s Watson supercomputer recommended «unsafe and incorrect» cancer treatments, internal documents show//STAT+.URL: https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf [↑](#footnote-ref-8)
9. Chiappa S. Path-Specidic Counterfactual Fairness // [Silvya Chiappa.] URL: https://csilviavr.github.io/assets/publications/silvia19path.pdf [↑](#footnote-ref-9)
10. Vincent J. Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women // The Verge. URL: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report Hamilton I. A. Why It’s Totally Unsurprising That Amazon’s Recruitment AI Was Biased against Women // Business Insider. URL: <https://www.businessinsider.in/why-its-totally-unsurprising-that-amazons-recruitment-ai-was-biased-against> women/articleshow/66192889.cms [↑](#footnote-ref-10)
11. Strickland E. Racial Bias Found in Algorithms That Determine Health Care for Millions of Patients // IEEE Speсtrum. URL: https://spectrum.ieee.org/the-human-os/biomedical/ethics/racial-bias-found-in-algorithms-that-determine-health-care-for-millions-of-patients [↑](#footnote-ref-11)
12. RobinetteP., HowardA., Wagner A.R. Conceptualizing overtrust in robots: Why do people trust a robot that previously failed? // Autonomy and Artificial Intelligence: A Threat or Savior? Cham: Springer International Publishing. 2017. P. 129–155. DOI: https://link.springer.com/chapter/10.1007/978-3-319-59719-5\_6; Wagner A. R.,Borenstein J., Howard A.Overtrust in the Robotic Age // Communications of the ACM. 2018. Vol. 61, no 9. P. 22–24. DOI: https://dl.acm.org/doi/fullHtml/10.1145/3241365 [↑](#footnote-ref-12)