AEQUITAS

D2.1 Доклад за видовете предубеденост при ИИ

Съфинансирано от Европейския съюз

AEQUITAS

Съдържание

Въведение 3

Медицински приложения с ИИ 5

Етика и предубеденост в медицината и ИИ 6

Биоетика и предубеденост в медицината 6

Етиката при ИИ и предубедеността 9

Предубедеността при системите с ИИ 12

Предварително съществуващи видове предубеденост 12

Казус: Диагностициране на сърдечно-съдови заболявания при жените 12

Технически видове предубеденост 13

Казус: Точност на моделите за прогнозиране на риска от инсулт при чернокожо и бяло население 13

Възникващи видове предубеденост 14

Казус: промени в набора от данни 14

Видове предубеденост, специфични за процеса на машинно обучение/ИИ 15

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

Представителна предубеденост 19

Измервателна предубеденост 20

Агрегационна предубеденост 22

Предубеденост при обучението 24

Предубеденост при оценката 25

Предубеденост при внедряването 27

Въздействия върху политиките 27

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

Въведение

Част от проекта Aequitas е създаване на база данни за полови и расови видове предубеденост при медицинските приложения с изкуствен интелект (ИИ), по специално фокусирани върху три заболявания: сърдечно-съдови заболявания, диабет и депресия.

За да изпълнят тази задача, партньорите в консорциума първо трябва да съберат различни източници на информация за посочените по-горе видове предубеденост. UKK, като ръководещ задачата и експерт в областта, организира дейността по събиране на информацията и предостави шаблона за картографиране, който партньорите използваха, за да картографират източниците, което гарантира лесното прехвърляне на нужната информация в базата данни.

Настоящият доклад представя теоретичните и научните основи, които са послужили като ръководство при избора на дейността по събиране на данни и шаблона за картографиране, придружени от казуси, показващи различните видове предубеденост; политическите последици, които различните видове предубеденост, предизвикани от биомедицинския ИИ, оказват върху правата, защитени от Хартата на основните права на ЕС; описание на дейността по събиране на данни; шаблона за картографиране; списък с източници, събрани от партньорите по AEQUITAS; и други подкрепящи материали. Останалата част от доклада е структурирана както следва:

Първо, представяме историята относно нашата работа чрез въведение за медицинските приложения с ИИ, и понятието за предубеденост в компютърните системи и медицината. Започваме с фокус върху медицината, като първо представяме как се проявяват расовите и половите видове предубеденост в медицинското обслужване, а след това как моралните проблеми, възникващи в практиката на медицината и биомедицинските изследвания, се разглеждат от биоетиката, като предлагаме кратко въведение в четирите принципа на биоетиката (автономност, ненанасяне на вреда, хуманност и справедливост).

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

После, преминаваме към областта на ИИ като представяме видовете предубеденост, които могат да бъдат конкретно наблюдавани в системите с изкуствен интелект, тъй както се проявяват в процеса на машинно обучение и изкуствен интелект. Всеки вид предубеденост е придружен от примери и казус за предубежденията, основани на пола и расата, и тяхното въздействие на обществено равнище по отношение на трите основни заболявания, които са обект на проекта AEQUITAS (сърдечно-съдови заболявания, диабет и депресия), извлечени от източници, събрани от партньорите по AEQUITAS след приключването на T2.2. Когато това не беше възможно, защото събраният материал не показваше ясно конкретния вид предубеденост на ИИ, беше представен алтернативен казус от друга медицинска област, който лесно можеше да се обобщи за целевите заболявания на AEQUITAS. Описанията на казусите, заедно със събраните източници, се основават на допълнителни научни ресурси, необходими за тяхното подкрепа.

Накрая, в раздела „Въздействия върху политиките“ се демонстрира как различните видове AI предубеденост засягат основните права, защитени от Хартата на ЕС, най-вече принципите на човешкото достойнство, равенството пред закона, недискриминацията, както и правото на лична неприкосновеност, правото на здравеопазване, защита на данните и правото на ефективна правна защита, като се стига до заключението, че могат да бъдат въведени предпазни мерки при оценката на съответствието, мониторинга след пускането на пазара и обществените поръчки.

Докладът завършва с библиография и следните приложения:

Приложение 1: Метод за събиране и картографиране на източници, който съдържа шаблона за картографиране и описва процеса на събиране, картографиране и оценка на информацията, проведен по време на задачите T2.1 и T2.2.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Приложение 2: Съдържа подкрепящи материали за задачите T2.1 и T2.2, т.е. слайдове от срещите на партньорите, описващи процеса, представени от UKK.

Приложение 3: Списък на източниците, събрани от партньорите на AEQUITAS.

Медицински приложения с ИИ

Възходът на приложенията с ИИ през последните години силно повлия медицината, включително цифровизиране на събирането на данни, машинно обучение и изчислителна инфраструктура (Yu et al., 2018). Конкретно навлизането на алгоритмите за дълбоко обучение в сфери като компютърното зрение и обработването на естествения език революционизираха компютърните приложения в радиологията, патологията, кардиологията, диабетологията, психиатрията, онкологията, т.н. (Esteva et al., 2019; Koteluk et al., 2021; Rajpurkar et al., 2022; Gou et al., 2024). Световната здравна организация (СЗО) изброява следните области на приложение на системите с ИИ в здравеопазването: диагностика и диагностика на базата на прогнози, клинични грижи, научни изследвания и разработване на лекарства, управление и планиране на здравните системи, обществено здраве и наблюдение на общественото здраве, насърчаване на здравето, профилактика на заболявания, наблюдение на базата на прогнози, готовност за извънредни ситуации и реагиране при епидемии (Световна здравна организация, 2021).

Въпреки това, навлизането на приложенията с изкуствен интелект в медицината е свързано с редица предизвикателства, като например предизвикателства при внедряването, включително доверие в моделите и ограничения на данните, въпроси, свързани с отчетността, които включват регулаторни предизвикателства и правилно разпределение на отговорностите, както и осигуряване на равнопоставеност чрез етично използване на данните, справедливо разпределение на ползите и откриване и намаляване на предубедеността (Rajpurkar et al., 2022).

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

Проектът AEQUITAS е насочен към случаите на полова и расова предубеденост при сърдечно-съдовите заболявания, диабета и депресията. Медицинските приложения с ИИ подпомагат сърдечно-съдовата здравна грижа чрез помощ при вземането на клинични решения, телемедицината и оценката на риска, персонализираното лечение, прогнозната аналитика и дистанционното наблюдение (Bernstein et al., 2025; Naskar et al., 2025), подобряват контрола на диабета (включително наблюдението на пациентите и самоконтрола), диагностиката, лечението и профилактиката (Contreras & Vehi, 2018; Khalifa & Albadawy, 2024; Naskar et al., 2025; Sheng et al., 2024), докато по отношение на депресията, те участват в скрининга, диагностиката и лечението (Alhuwaydi, 2024) с особен акцент върху откриването и скрининга с използването на големи езикови модели LLMs (Cao et al., 2025; Kumari et al., 2025; Mao et al., 2023; Wang et al., 2025). Във всички горепосочени области съществуват предизвикателства, свързани с предубеденост, например по отношение на сърдечно-съдовите заболявания (van Assen et al., 2024), диабета (Cronjé et al., 2023) и депресията (Dang et al., 2024).

Предизвикателствата, като например предубедеността, независимо дали в медицината или в изкуствения интелект, се разглеждат чрез комбинация от биоетика и етика при изкуствения интелигент. В следващата част представяме кратък преглед на тези два вида приложни етични области, които послужиха като теоретична и научна основа за разработването на шаблона за картографиране на предубедеността.

Етика и предубеденост в медицината и ИИ

Биоетика и предубеденост в медицината

Предубедеността в медицината е добре документирана; например Hammond et al. (2021) разглеждат когнитивната предубеденост, която представлява системни грешки в мисленето поради ограничения на човешката способност за обработка или неподходящи ментални модели. FitzGerald & Hurst (2017) изучават имплицитната предубеденост, включваща асоциации извън съзнателното

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

възприятие, които водят до негативна оценка на дадено лице въз основа на несъществени характеристики като раса или пол.

Расовата предубеденост в медицината е добре проучена в случая на САЩ, например, където е документирано, че афроамериканците, както и тези от други малцинствени групи, получават по-малко процедури и по-нискокачествено медицинско обслужване, като им се предоставя по-малко агресивно лечение, по-ниски проценти на хирургично лечение и по-малко препратки към специалисти в сравнение с белите лица (Bowser, 2001; Williams & Wyatt, 2015).

Половата предубеденост може да се дължи на полова слепота и стереотипни предразсъдъци за мъжете и жените (Hamberg, 2008), в допълнение към общата липса на знания за функционирането на женското тяло и биологичните му разлики от мъжкото тяло. Например, жените на възраст 50 години и повече в критично състояние са по-малко вероятно да бъдат приети в интензивно отделение в сравнение с мъжете в критично състояние (ICU) (Bierman, 2007), дори мъжките мишки модели са по-често използвани от женските модели в основните, предклиничните и хирургичните биомедицински изследвания (Yoon et al., 2014).

Важно е също да се отбележи, че ЛГБТ+ лицата са обект на дискриминация по отношение на достъпа до здравни грижи и са подложени на стереотипи, които не засягат хетеросексуалното население. Тези социални и културни фактори затвърждават дискриминацията и оказват влияние върху здравето. Например, проучване в САЩ, базирано на данни от Националното проучване за здравето (NHIS) за 2013–2014 г., установи, че възрастните ЛГБ лица съобщават за по-високи нива на лошо здраве, функционални ограничения, тежки психологически проблеми и затруднения при достъпа до здравни грижи в сравнение с хетеросексуалните си съграждани. Тези неравенства се дължат на стреса, на който са подложени малцинствените групи, и на многостранната социална маргинализация (Liu et al., 2023).

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

От друга страна, медицината като дисциплина се придържа към високи етични стандарти от древността до наши дни (Baker & McCullough, 2008). В продължение на векове се е очаквало, че лекарят ще следва етични правила на професионална отговорност, установени от стандартите на тяхната професия, както са прокламирани чрез професионални норми като се почне от Хипократовата клетва от 400 г. пр. н. е. (Miles, 2005), до Женевската и Хелзинкската декларации (Tröhler, 2008). Както е посочено от Vevaina et al. (1993), лекарите носят отговорност да спазват етичния кодекс на своята професия поради инвестицията, която обществото прави в тяхното образование (финансова и използването на своите членове като учебни материали по време на обучението и кариерата на лекарите), както и поради фактическия монопол, който се предоставя на тяхната професия чрез лицензирането.

Биомедицинската етика (или биоетиката) е област от практическата (или приложната) етика, която се занимава с моралните проблеми, възникващи в практиката на медицината и биомедицинските изследвания (Vevaina et al., 1993). Централни за биомедицинската етика са четирите принципа, определени от Beauchamp & Childress (2019):

1. Автономност: зачитане на способността за вземане на решения на автономните лица. Две общи условия са от съществено значение за автономността: свобода, изразяваща се в независимост от контролиращи влияния, и способност за действие, а именно способността за целенасочено действие.

2. Ненанасяне на вреда: избягване на причиняването на вреда.

3. Хуманност: предприемане на положителни стъпки, за да се помогне на другите, по-специално предотвратяване на злото или вредата, премахване на злото или вредата и насърчаване на доброто.

4. Справедливост: справедливо разпределение на ползите, рисковете и разходите. Справедливостта се тълкува като справедливо, равностойно и

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

подходящо отношение към отделните лица и групи, като се имат предвид многобройните неравенства в здравеопазването и научните изследвания, основани на раса, етническа принадлежност, пол и социален статус.

Етиката при ИИ и предубедеността

Въвеждането на изкуствен интелект и бързото развитие на приложенията с изкуствен интелект повдигнаха редица етични въпроси (Christoforaki & Beyan, 2022), сред които на преден план изпъкват предубедеността и дискриминацията.

Така етиката при ИИ се разви като област на практическа (или приложна) етика, която обхваща „набор от ценности, принципи и техники, които прилагат широко възприети стандарти за правилно и погрешно за направляване на моралното поведение при развитието и употребата на технологиите с ИИ“ (Leslie, 2019, p. 3).

Етиката при изкуствения интелект се основава както на биоетиката (четирите принципа, представени по-горе), така и на дискурса за правата на човека, като последният включва, наред с другото, правото на равна свобода и достойнство пред закона, защитата на гражданските, политическите и социалните права, всеобщото признаване на личността и правото на свободно и необременено участие в живота на общността (Leslie, 2019).

Четирите принципа на биоетиката, допълнени с още един - обяснимост (дефинирана като разбиране и отчетност на процесите на вземане на решения от ИИ), са представени за ИИ от Floridi et al. (2018) както следва:

1. Автономност, като способността на хората да решават дали да вземат решение, и съдържаща риска от прекалено делегиране на машините.

2. Ненанасяне на вреда, като предотвратяване на вреди, произтичащи от намеренията на хора или от непредвидимото поведение на машини.

3. Обществено и екологично благополучие, като насърчаване на благоденствието, запазване на достойнството и поддържане на планетата.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

4. Справедливост, като предотвратяване и премахване на вече съществуващи несправедливи дискриминации, както и на нови вреди, и осигуряване на равномерно разпределение на ползите от ИИ.

5. Обяснимост, определена като разбиране и отчетност на процесите на вземане на решения от ИИ.

По този начин етиката при изкуствения интелигент също се е фокусирала върху набор от принципи, базирани на четирите класически принципа на медицинската етика, както и върху други подходи, обобщени от Christoforaki & Beyan (2022). Въпреки това, както се отбелязва от Mittelstadt (2019), в сравнение с медицината, развитието на ИИ не разполага с (1) общи цели и правни задължения, (2) професионална история и норми, (3) доказани методи за превръщане на принципите в практика и (4) солидни правни и професионални механизми за отчетност, което подкопава успеха на принципния подход. Естествено, съществува и сложна регулаторна среда, която урежда развитието и използването на ИИ в ЕС, включително закони срещу дискриминацията, но тази тема е извън обхвата на настоящия доклад.

По отношение на човешките права според доклад от 2018 г., финансиран от Съвета на Европа (Комитета от експерти по интернет посредниците (MSI-NET), 2018), човешките права, които са особено засегнати от алгоритмите и от техниките за автоматизирано обработване на данни, включват:

- Свободно съдебно производство и справедлив процес
- Защита на личния живот и личните данни
- Свобода на изразяване
- Ефективни средства за защита
- Свобода на събранията и сдружаването

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

- Забрана на дискриминацията
- Социални права и достъп до обществени услуги
- Право на свободни избори

Предубедените алгоритми са изрично споменати като възможни фактори за дискриминация срещу социални групи въз основа на възраст, сексуална ориентация, раса, пол или социално-икономически статус (Комитета от експерти по интернет посредниците (MSI-NET), 2018, p. 27). Освен това, Рамковата конвенция на Съвета на Европа за изкуствения интелект и правата на човека, демокрацията и върховенството на закона изрично посочва, че държавите членки „приемат или поддържат мерки с цел да гарантират, че дейностите в рамките на жизнения цикъл на системите с изкуствен интелект зачитат равенството, включително равенството между половете, и забраната на дискриминацията, както е предвидено в приложимото международно и национално право“ (Рамкова конвенция на Съвета на Европа за изкуствения интелект и правата на човека, демокрацията и върховенството на закона, 2024 г., p. 4).

Гражданските организации като заинтересовани страни в екосистемата на здравеопазването (Vayena et al., 2018) могат да играят значителна роля в идентифицирането и справянето с предубедеността при ИИ и управлението на ИИ като цяло чрез застъпничество за етично развитие на ИИ, поемане на отговорност от заинтересованите страни, образование на обществеността, представляване на маргинализирани общности, формиране на политически и регулаторни рамки и насърчаване на сътрудничеството между правителства, технологични компании и обществеността (Korir, 2024).

В този теоретичен контекст са разработени изрично различни технически решения по отношение на предубедеността. В следващия раздел представяме класификация на видовете предубеденост, предизвикани от ИИ, която послужи като основа за нашия шаблон за картографиране, като се фокусираме върху тяхното въздействие върху дискриминацията по пол и раса. Човешките психически

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

предубеждения (Hofmann, 2023), например когнитивните предубеждения, като предубежденията за потвърждение или наличност, макар и с голямо влияние в медицината, се считат за извън обхвата на настоящия проект.

Предубедеността при системите с ИИ

Предубедеността в компютърните системи е дефинирана от Friedman & Nissenbaum (1996, p. 332) като термин „ [отнасящ се] до компютърни системи, които систематично и неправомерно дискриминират определени лица или групи от лица в полза на други. Системата дискриминира неправомерно, ако тя отказва възможност или благо или определя неблагоприятен изход на лице или група лица на основания, които са необосновани или неподходящи“.

Според Friedman & Nissenbaum (1996), видовете предубеденост в компютърните системи могат да бъдат разделени на три категории: предварително съществуващи видове предубеденост, технически видове предубеденост и възникващи видове предубеденост. В следващия подраздел разглеждаме всеки вид предубеденост и го илюстрираме с казуси, както са отразени в научната литература.

Предварително съществуващи видове предубеденост

Предварително съществуващите видове предубеденост произтичат от вече съществуващи видове предубеденост в социалните институции, практики и нагласи и съществуват независимо, обикновено преди създаването на системата. Този вид предубеденост се включват в системата съзнателно или несъзнателно, понякога дори когато създателите на системата се опитват да ги избегнат.

Казус: Диагностициране на сърдечно-съдови заболявания при жените

Сърдечно-съдовите заболявания (ССЗ) обикновено се възприемат като „мъжки болести“, което допринася за недооценяване на диагнозата и лечението при жените. Както е показано от Al Hamid et al. (2024), систематичен преглед на въпроса, ССЗ са по-рядко диагностицирани при жени, които или показват по-леки

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

симптоми от мъжете, или симптомите им са погрешно диагностицирани като стомашно-чревни или свързани с тревожност; по този начин на жените се предлагат по-малко диагностични тестове, лекарства и по-рядко се насочват към кардиолози и/или хоспитализация. Освен това, ако са били хоспитализирани, жените са имали по-малка вероятност да получат коронарна интервенция. Следователно рисковите фактори при жените са били подценявани от лекарите, особено от мъжете лекари. Предвид факта, че жените остават недостатъчно представени в областта на кардиологията (Fatunde et al., 2025), може да се заключи, че жените имат по-малка вероятност да получат подходящи здравни грижи поради вече съществуващата предубеденост.

Системите с ИИ се обучават с помощта на данни, събрани от съществуващите практики, така че система за диагностика на сърдечно-съдови заболявания, базирана на ИИ, ще включва тази предубедена нагласа, създавайки дискриминация срещу жените, независимо от избора при техническото внедряване.

Технически видове предубеденост

Техническите видове предубеденост възникват от технически ограничения или технически съображения, главно когато създателите на системата се опитват да направят човешките конструкции приложими за компютрите, като например количествено измерване на качественото, дискретизиране на непрекъснатото или формализиране на неформалното. Освен това, изваждането на алгоритмите от контекста на средата, в която работят, може да доведе до това, че те да не третират всички групи справедливо при всички значими условия.

Казус: Точност на моделите за прогнозиране на риска от инсулт при чернокожо и бяло население

Hong et al. (2023) проведоха ретроспективно проучване на прогнозната точност на риска от инсулт, като сравниха съществуващите модели за прогнозиране на риска

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

от инсулт и нови техники за машинно обучение, включващи, наред с други критерии, расата на пациентите. Всички алгоритми показаха по-лоша дискриминация при чернокожи лица, отколкото при бели лица. Според авторите тази ситуация може да се дължи на рискови фактори, които не са отразени в данните, като вид застраховка, езикови бариери и други фактори, произтичащи от различен достъп до здравни услуги, т.е. данните са извадени от контекста на социално-икономическата среда, в която са били генерирани. В същото време всички гореспоменати рискови фактори са конструкти, които е трудно да бъдат представени във форма, подходяща за компютри. Към всичко това можем да добавим, че най-съвременните алгоритми за изкуствен интелект по своята същност са непрозрачни по отношение на характеристиките, които избират, за да постигнат висока точност (Knight, 2017), което прави дори техните създатели неспособни да обяснят как работят и по този начин да контролират дали някой от гореспоменатите социално-икономически фактори наистина се взема под внимание във вътрешното функциониране на системата с изкуствен интелект.

Възникващи видове предубеденост

Възникващите видове предубеденост се проявяват в контекста на използване с реални потребители, обикновено след като проектът е завършен, в резултат на променящите се обществени знания, които не могат да бъдат или не са включени в проектирането на системата, или в резултат на популация с различни познания или културни ценности от предполаганите в проектирането.

Казус: промени в набора от данни

Промените в набора от данни представляват несъответствие между разпределението на наборите от данни за обучение и тестване по време на разработването на алгоритъма и могат да доведат до различна ефективност на ниво подгрупа (Chen et al., 2023).

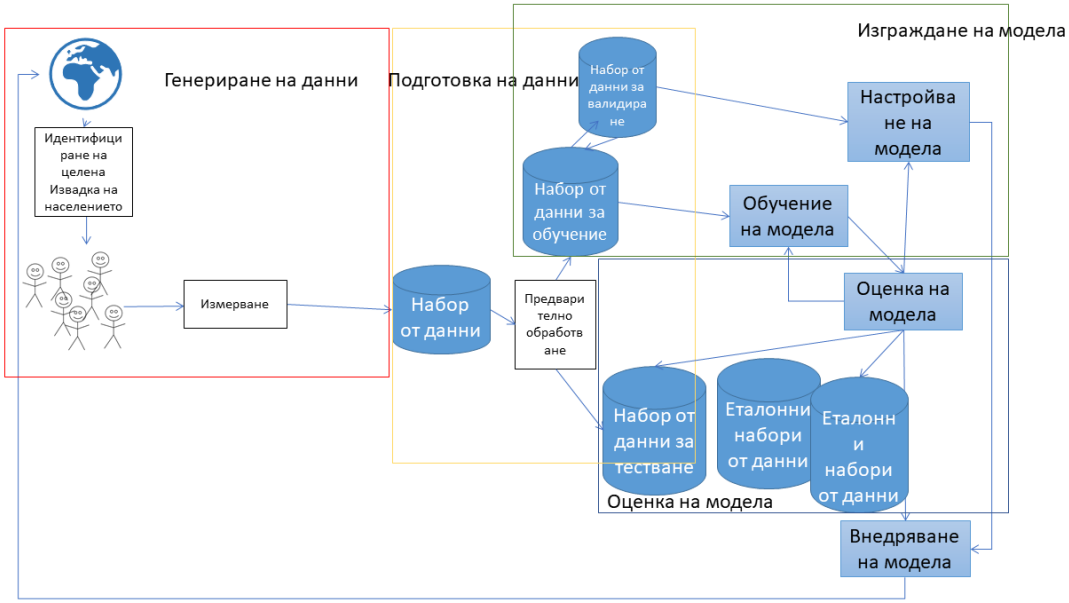
Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

При откриването на рак на кожата, например, много от наборите от данни за образни изследвания, използвани за обучение на алгоритми за изкуствен интелект за откриване на рак на кожата, са с произход от страни с население с по-светла кожа (Guo et al., 2021), което води до недостатъчно представяне на определени демографски групи. Алгоритмите с ИИ, обучени с тези набори от данни, имат по-ниска ефективност, когато се прилагат в страни с по-разнообразно население, като дискриминират хората с тъмна кожа. Наборите от данни са трудни и скъпи за събиране, анотиране и валидиране, което налага системите с ИИ, разработени в страни с ниски и средни доходи, да разчитат на обществено достъпни набори от данни, които може да не отразяват разпределението на населението им, което води до несъответствие между изходната и целевата популация. Същото може да се случи и в страни с високи доходи, например поради промени в населението в резултат на увеличената имиграция или поради различия в самоопределянето на расата. Както е отбелязано от Chen et al. (2023), „тъй като вече е прието, че расата е социална конструкция и че има по-голяма генетична вариабилност в рамките на дадена раса, отколкото между расите“ […] „медицинската общност започна да осъзнава, че таксономиите от миналото не представят адекватно групите хора, които претендират да представляват“ и „могат да замъглят културата, историята, социално-икономическия статус и други фактори, които влияят на равнопоставеността“.

Видове предубеденост, специфични за процеса на машинно обучение/ИИ

Макар горното да важи за всички компютърни системи, приложенията с изкуствен интелект имат по-специфични изисквания, затова се нуждаехме от по-подробна таксономия. Вследствие на това решихме да следваме класификацията на видовете предубеденост, представена от Suresh & Guttag (2020), тъй като тя идентифицира видовете предубеденост на всеки етап от машинното обучение/изкуствения интелект, както е илюстрирано на фигура 1.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI



Фигура 1 Процес на машинно обучение/ изкуствен интелект. Изображение, адаптирано от Suresh & Guttag (2020).

Типичен процес на машинно обучение/ изкуствен интелект може да бъде описан по следния начин:

- **Генериране на данни**. Създаването на система с машинно обучение/ изкуствен интелект започва с генерирането на данни. Това включва първо събиране и подготовка на данни за съставяне на набор от данни за системата с изкуствен интелект. Съществуващите данни в света трябва да бъдат събрани чрез идентифициране на целева извадка от населението. Следващата стъпка е да се определят и измерят характеристиките, които са от значение за приложението, което ще се внедрява, и/или да се анотират данните с подходящи етикети. Това е скъп и продължителен процес, така че в повечето случаи практикуващите ИИ използват съществуващи набори от данни (публични или закупени).

- **Подготовка на данни**. На този етап може да се наложи предварителна обработка на данните (например почистване, нормализиране и обработка на липсващи стойности). На този етап наборът от данни се разделя на три части, а именно *набор от данни за обучение*, действителният набор от данни, използван за обучение на модела; *набор от данни за валидиране*, извадка от

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

данни, използвана за оценка на пригодността на модела към набора от данни за обучение, докато се настройват хиперпараметрите на модела (конфигурация на модела, която не може да бъде научена от данните, например броят на слоевете и невроните в модел на невронна мрежа); и *набор от данни за тестване*, част от данните, използвани за оценка на крайния модел, предоставящ златен стандарт, след като моделът е напълно обучен.

- **Изграждане на модел**. На този етап моделът се обучава на базата на обучителните данни и се фино настройва чрез коригиране на хиперпараметрите на набора от данни за валидиране.

- **Оценка на модела**. Обученият модел се оценява с помощта на набора от данни за тестване и понякога с помощта на еталонни набори от данни, които са независимо съставени набори от данни, използвани за демонстриране на устойчивостта на модела и/или за сравнение с други методи.

- **Внедряване на модела**. Прилагане на модела в реална среда. Това може да доведе до промени в зависимост от резултатите и може да създаде обратна връзка с началото на процеса.

Като вземем предвид гореописаните фази на процеса на машинно обучение/изкуствен интелект (Suresh & Guttag, 2020) и идентифицираме следните категории предубеденост: историческа, представителна, измервателна, агрегационна, предубеденост при обучението, предубеденост при оценката и предубеденост при внедряването. В следващите подраздели ще дефинираме гореизброените видове предубеденост и ще предложим казуси от източниците, събрани за проекта.

Историческа предубеденост

Историческата предубеденост съответства на предварително съществуващата предубеденост, както е дефинирана от Friedman & Nissenbaum (1996), която включва вече съществуващи предразсъдъци и стереотипи в данните. Пример за

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

това може да се види в анализа на Calderone (1990), който изследва дали честотата на болка и седативни лекарства, прилагани на пациенти след операция за коронарен байпас, се различава в зависимост от пола и възрастта на пациента. Резултатите показват, че на мъжете пациенти и пациентите на възраст 61 години или по-млади са били прилагани обезболяващи лекарства значително по-често, отколкото на жените пациенти и пациентите на възраст 62 години и повече, на които вместо това са били прилагани успокоителни лекарства значително по-често. Казусът за точността на моделите за прогнозиране на риска от инсулт сред чернокожото и бялото население е представен в подраздела „Предварително съществуващи предубеждения“ „Предварително съществуващи предубеждения“. Ние обаче ще представим друг казус, който показва историческите предубеждения по отношение на използването на изкуствен интелект в областта на психичното здраве.

Казус: Изкуствен интелект в областта на психичното здраве и видовете предубеденост на езиковите модели

Straw & Callison-Burch (2020) представят систематичен преглед на литературата за използването на NLP в психичното здраве с цел да се идентифицира как тези видове предубеденост могат да увеличат неравенствата в здравеопазването. Моделите с ИИ, които използват NLP за профилиране на психичното здраве, събират големи масиви от данни за изразителен език, обикновено придобити от социални медии, онлайн форуми, блогове и чат стаи. Тези данни обаче вече са повлияни от личния произход и социалния контекст на дадено лице.

По-конкретно, по отношение на пола и езика, има обширна библиография (за английския език), обобщена от Pennebaker et al. (2003), която разкрива разлики в употребата на думи от жените и мъжете. Например, жените използват по-малко категорична реч, което се проявява в по-голяма учтивост, по-малко псувни, повече усилващи думи (напр. наистина, така) и повече уклончиви изрази (т.е. квалификатори или несигурни думи като нещо като, може би или може би). Мъжете, от друга страна, са описани като директивни, прецизни и по-малко

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

емоционални в използването на езика, което се характеризира с позоваванията на количество, съдителни прилагателни (например, добър, глупав), елиптични изречения („Страхотна снимка.“) и позоваванията на „аз“. Както отбелязват авторите, тези разлики са в съответствие със социологическата рамка на половите различия, но могат да бъдат приписани и на алтернативни обяснения, като по-голямата социална ангажираност на жените.

По отношение на психичното здраве, мъжете и жените пишат различни предсмъртни писма, в които изразяват своята суицидна тревога; жените вътрешно потискат негативните емоции, докато мъжете изразяват нарастващ гняв (Straw & Callison-Burch, 2020). Система с изкуствен интелект, която проверява психичното здраве на единия пол, може да е неподходяща за другия (а това е при разглеждане на пола в бинарен контекст, който изключва голяма част от населението).

Представителна предубеденост

Представителната предубеденост възниква, когато извадката за разработване не отразява адекватно част от населението по време на фазата на събиране на данни. Това може да възникне по следните начини: при определяне на целевото население, ако то не отразява използващото население; при определяне на целевото население, ако то съдържа недостатъчно представени групи; при вземане на извадка от целевото население, ако методът на вземане на извадка е ограничен или неравномерен. Представителната предубеденост води до лоша генерализация за подгрупа от потребителското население. Типичен пример за представителна пристрастност е откриването на рак на кожата, тъй като много набори от данни за изображения не представят адекватно определени демографски групи, което води до обучение на моделите за машинно обучение върху изображения предимно на хора с по-светла кожа (Guo et al., 2021). Като вземаме предвид целевите заболявания на проекта AEQUITAS, представяме казус за представителната предубеденост по отношение на расата при диабет тип 2.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

Казус: Оценка на расовата предубеденост в алгоритмите за прогнозиране на риска от диабет тип 2

Според Cronjé et al. (2023), по отношение на населението на САЩ, въпреки сравнително по-ниския риск, неиспано-американските бели групи остават надпредставени в литературата за прогнозиране на риска от диабет. В друг преглед на етнорасовата равнопоставеност при изкуствения интелигент за управление на диабета, в прегледаните статии, които съобщават за расата, средното разпределение е 69,5% бели, 17,1% чернокожи и 3,7% азиатци, докато само 2 статии съобщават за включването на участници от коренното американско население (Pham et al., 2021).

Добре документирано е, че неравенствата в резултатите от диабета се дължат до голяма степен на сложни, взаимосвързани социални детерминанти на здравето, включително достъп до здравословна храна, качествено здравно обслужване, застрахователен статус, образователни бариери и разлики в степента на внедряване на технологии. Тези резултати включват по-високи нива на усложнения и по-лош гликемичен контрол сред малцинствените и нискодоходните групи от населението (Alipour & Alipour, 2025).

В резултат на това система с ИИ, обучена на съществуващи набори от данни, би генерализирала погрешно, което би довело до предубедени прогнозни модели, които могат да благоприятстват лица от определени расови групи, например при превантивни действия.

Измервателна предубеденост

Измервателната предубеденост възниква при избора, събирането или изчисляването на характеристики и етикети, които да се използват в проблем с прогнозиране, особено при използване на прокси (приближение на конструкция, която не е директно кодирана или наблюдаема). Могат да възникнат следните проблеми: използване на прокси, което прекалено опростява по-сложна конструкция, а методите за измерване и точността варират между групите. Пример

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

за този вид отклонение може да бъде намерен в анализа на Obermeyer et al. (2019), където разходите за здравеопазване са били използвани като прокси за прогнозиране и класифициране на пациентите, които биха имали най-голяма полза от допълнителни грижи, което е довело до расова дискриминация. Въпреки това, разходите за здравеопазване са лош показател за здравните нужди, тъй като чернокожите пациенти, които са изправени пред несъразмерни нива на бедност, често харчат по-малко за здравни грижи от белите. Поради тази предубедена нагласа алгоритъмът погрешно стига до заключението, че чернокожите са по-здрави от белите пациенти с еднакви заболявания, като по този начин ги класифицира като пациенти с по-ниска приоритетност при достъпа до здравни услуги.

Други източници на предубеденост в измерването могат да възникнат, когато методът на измерване варира между групите, например когато две групи се наблюдават за едно и също поведение, но едната от тях се наблюдава по-стриктно или по-често от другата. По същия начин точността на измерването може да варира между групите, което в медицинските приложения може да доведе до систематично по-високи нива на погрешна диагноза или недооценена диагноза в определени групи. Например, лекарите са по-склонни да подценяват болката на чернокожите пациенти в сравнение с нечернокожите пациенти поради погрешни вярвания за биологичните разлики между чернокожи и бели, което води до по-малка вероятност чернокожите пациенти да получат обезболяващи лекарства, а ако им бъдат дадени, те получават по-малки количества (Hoffman et al., 2016).

Казус: Расови и етнически различия във връзката между средната стойност на глюкозата и хемоглобина A1c

Тестът A1C измерва средното количество глюкоза (захар) в кръвта и се използва за откриване на предиабет или за диагностициране на диабет тип 2. Въпреки това, A1C е само индиректна мярка и не е причинно свързана със здравните резултати, тъй като има много начини, по които връзката между директните измервания на гликемията (концентрацията на глюкоза в кръвта) и A1C може да бъде пряко

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

променена. Съществува дори значителна вариация във връзката между гликемията и A1C между отделните индивиди и дори в рамките на един и същ индивид във времето. Освен това, проучвания са отчели значително по-високи нива на хемоглобин A1c (A1C) при афроамерикански пациенти в сравнение с бели пациенти със същата средна глюкоза (Karter et al., 2023).

Ако система с ИИ, предназначена за диагностициране на диабет, е обучена да използва резултатите от A1C теста като заместител на гликемията, без да отчита други фактори, като расата на пациента, това може да доведе до преждевременна диагноза на диабет и неподходящо лечение, което да доведе до предубеденост в качеството на здравните грижи и неравенства в здравеопазването. Въпреки това, както е отбелязано от Alipour & Alipour (2025), систематичен преглед на предубедеността, която може да повлияе на равнопоставеността на AI/ML моделите при диабет (включително предубеденост при измерването), макар че проучените изследвания изрично споменават, че предубедеността при измерването може да се разпространи чрез моделите с ИИ, ако не бъде коригирана, никое от тях не е отчело такава предубеденост по време на разработването на модела, не я е коригирало изрично или не е докладвало за коригиране на разликите в точността на измерването.

Агрегационна предубеденост

Агрегационната предубеденост възниква, когато се използва универсален модел за набор от данни, който включва разнообразни групи от хора или неща.

Можем да разгледаме примера с картографиране на входни данни (например доходът на дадено лице) към етикети, които ги описват (например нисък, среден, висок), като се приема, че те са еднакви за всички подгрупи от данните. В действителност произходът или културата на дадено лице могат да променят действителното значение на тези цифри. Например, „висок“ доход в малко селско градче или в страна с ниски или средни доходи може да означава нещо много по-различно отколкото в голям град или в страна с високи доходи.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Казус: Цифрови инструменти за пасивно наблюдение на депресията

Използването на цифрови инструменти за измерване на физиологични и поведенчески променливи за пасивно наблюдение на депресията е разгледано от De Angel et al. (2022), систематичен преглед на въпроса. В прегледаните статии са проучени връзките между депресията и обективните поведенчески данни, получени от сензори на смартфони и носими устройства. Тези данни са картографирани в характеристики, използвани от моделите на изкуствен интелект за правене на прогнози, съответстващи на сън, физическа активност, циркаден ритъм, общителност, местоположение и използване на телефона.

Въпреки това авторите подчертават хетерогенността, която възниква от разнообразието на методите, използвани за създаване на тези характеристики. Например, характеристиката „качество на съня“ може да бъде определена чрез измерване на броя на събужданията, общия брой минути в състояние на будност или съотношението между събуждане и сън в една сесия на сън, като трябва да се вземат предвид и разликите в начина, по който сензорите в различните устройства описват едно събитие като „сън“. Тъй като всички горепосочени различия не се вземат под внимание и се обединяват под общото наименование „качество на съня“, а също така, тъй като наборът от данни може да бъде получен от хора или групи с различен произход, култура или норми, тази характеристика може да има различно значение за всяка от тези групи или индивиди.

Обединяването на такива данни в една единствена характеристика може да доведе до система, която не е подходяща за никоя група или която дава предимство на доминиращото население, ако има и предубеденост в представянето. Например, има доказателства, че съществуват полови различия в съня между мъжете и жените, докато последните често са недостатъчно представени в изследванията на съня. Освен това, други фактори, които обикновено не се вземат под внимание при моделите и нарушенията на съня, не разграничават пола като социална конструкция от биологичния пол и не отчитат

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

интерсекционалните идентичности, определени от възраст, раса и социално-икономически клас (Lok et al., 2024).

Предубеденост при обучението

Предубеденост при обучението възниква, когато изборът на модели усилва разликите в представянето между различните примери в данните. Един пример за това е диференциалната поверителност – механизъм, използван в системите за изкуствен интелект, който гарантира, че чрез проверка на резултатите от системата не е възможно да се определи дали данните на конкретно лице са били включени в оригиналния набор от данни. Диференциалната поверителност се използва в набори от данни в областта на здравеопазването, за да се защити чувствителна информация за пациенти, например в случая на редки заболявания, при които всеки пациент е повече или по-малко уникален в ограничена област, обхванатата от болница, така че дори и данните да са анонимизирани, не е много трудно да се определи самоличността на лицето. Въпреки това е доказано, че диференциалната поверителност намалява влиянието на недостатъчно представените данни върху модела; следователно, ако системата с изкуствен интелект е предубедена от самото начало, прилагането на мярка за повишаване на поверителността засилва тази предубеденост още повече (Bagdasaryan & Shmatikov, 2019).

Казус: Диференциална поверителност и различия в здравеопазването

През септември 2018 г. Американската статистическа служба обяви, че ще въведе диференциална поверителност за продуктите, произведени от данните от преброяването на населението през 2020 г. Въпреки това, Santos-Lozada et al. (2020) проучиха как въвеждането на диференциална поверителност може да промени знанията за различията в смъртността, особено за расовите или етническите малцинства в малките райони и по-малко урбанизираните области. Резултатите им показват, че диференциалната поверителност ще повлияе по-силно на оценките за смъртността при неиспано-американските чернокожи и испано-американците, отколкото на оценките за неиспано-американските бели.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Тези констатации бяха подкрепени от Kurz et al. (2022), които показват, че прилагането на диференциална поверителност към същите данни може да доведе до погрешно представяне на процента на участие в програмата Medicaid сред вече маргинализираните расови и етнически групи. По-конкретно, тези проценти за определени комбинации от окръг, раса и етническа принадлежност се различаваха между резултатите от диференциалната поверителност и оригиналните данни, като понякога надвишаваха 10 %. Освен това, неиспано-говорящите бели лица бяха единствената етническа и расова подгрупа, за която алгоритъмът за диференциална поверителност точно отрази процентите на участие в Medicaid. Това заключение може да има важни последици за здравната политика, тъй като данните от преброяването се използват за планиране на правителствени програми, разпределяне на ресурси и оценка и проследяване на политиките.

Предубеденост при оценката

Предубеденост при оценката възниква, когато референтните данни, използвани за дадена задача, не представляват използващата ги популация. Референтните данни са стандартизирани набори от данни, използвани за измерване на качеството на даден модел, което позволява количествено сравнение на моделите. Впоследствие съществува риск от насърчаване на разработването и внедряването на модели, които функционират добре само върху подмножество от данните, представени в референтните данни. По този начин може да възникне дискриминация срещу уязвими подгрупи или лица, ако референтните данни са обект на историческа, представителна или измервателна предубеденост.

В здравеопазването причините за недостатъчното представяне на определени групи от населението в наборите от данни могат да бъдат или липсата на лица или групи в наборите от данни (например бременни жени поради етични ограничения), или неправилното или неподходящо категоризиране на хората в групи (например категории „смесена етническа принадлежност“ или „други“). Основните причини за това могат да включват социални и технически или правни/етични причини, като структурни бариери за получаване на здравни грижи,

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

технически пречки за събиране или дигитализиране на съответните здравни данни, индивидуални и структурни ограничения по отношение на съгласието за споделяне на данни и правни или етични ограничения за споделяне на данни, които пречат на достъпността на данните, между другото (Arora et al., 2023). В резултат на това системите с изкуствен интелект, калибрирани според такива еталони, могат да имат по-ниска ефективност, когато се прилагат към лица от недостатъчно представена група. Важно е обаче да се отбележи, че валидността на еталоните е по-общ проблем и не се ограничава до предубедеността (Brooks, 2025).

Казус: Набори от данни за изображения на кожата

Наборите от данни за изображения на кожата не представят адекватно определени демографски групи, тъй като повечето изображения в тези набори произхождат от популации в Северна Америка или Европа и предимно изобразяват хора със светла кожа (Guo et al., 2021). Поради високата цена и трудността при създаването на тези набори от данни, освен за обучение на модели, те могат да се използват и като еталони.

Казусът, който илюстрира възникващи видове предубеденост, т.е. наборите от данни с изображения на рак на кожата, използвани за обучение на модели за прогнозиране, е пример за неподходящ еталон, когато потребителската популация произхожда от недостатъчно представени групи (Guo et al., 2021). Подобен случай, макар и несвързан с ИИ, показва общата същност на проблема, който е свързан с пулсоксиметрите (устройства, които измерват насищането на кръвта с кислород, използвани например при сърдечен удар или сърдечна недостатъчност), които се оказват по-точни при светла кожа (Sjoding et al., 2020).

Представителността, измерването, агрегирането, ученето и оценката могат да бъдат отнесени към техническите видове предубеденост, дефинирани от (Friedman & Nissenbaum, 1996).

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

Предубеденост при внедряването

Предубедеността при внедряването възниква, когато има несъответствие между проблема, който моделът трябва да реши, и начина, по който той се използва в действителност, което може да причини вреда, особено когато се комбинира с когнитивни пристрастия като предубеждение за потвърждение и автоматизация. Предубедеността при внедряване е същата като възникващите видове предубеденост , дефинирани от Friedman & Nissenbaum (1996).

Казус: промяна на домейна

Случаят с промяна на данните е документиран в подраздела възникващи видове предубеденост относно откриването на рак на кожата. Освен това можем да дефинираме случая с промяна на домейна, който възниква, когато дадена система е внедрена, е получила регулаторно разрешение и се използва в клиничната практика, но се прилага към различна група пациенти от тази, за която е била обучена. Например, дадена система може да бъде разработена за болница в страна с висок доход и да бъде внедрена в страна с нисък или среден доход, без да се вземат предвид фактори като социодемографските характеристики на пациентите или дали пациентите имат същото общо ниво на риск в сравнение с тези, включени в данните за обучение (Vokinger et al., 2021).

Въздействия върху политиките

Доказателствата, представени в Доклад D2.1, показват, че видовете предубеденост, основани на пол и раса, в биомедицинската ИИ не са случайни или изолирани технически недостатъци, а системни рискове, които възникват през целия жизнен цикъл на системите с ИИ, използвани в здравеопазването. При сърдечно-съдовите заболявания, депресията и диабета видовете предубеденост произтичат от исторически изкривени клинични данни, неравностойни диагностични практики, прокси променливи, които кодират структурни

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

неравенства, и контексти на внедряване, които разпределят неравномерно както ползите, така и вредите. Тези констатации потвърждават, че биомедицинския ИИ засяга пряко множество права и принципи, защитени от Хартата на основните права на ЕС, най-вече принципите на човешкото достойнство, равенството пред закона и недискриминацията, както и правото на неприкосновеност на личността, правото на здравеопазване, защита на данните и правото на ефективна правна защита.

На този фон европейските и националните политически рамки, регулиращи ИИ в здравеопазването, трябва да третират намаляването на предубедеността не като доброволно етично допълнение, а като задължителен компонент на законното и съобразено с правата внедряване на ИИ. Европейските и националните регулаторни усилия по отношение на ИИ в здравеопазването трябва да се разглеждат като част от по-широката рамка на основните права, регулираща ИИ (вж. Novossiolova, 2025; Novossiolova et al., 2025; Kasapi, 2025). Законът на ЕС за ИИ осигурява необходимата регулаторна основа, като класифицира по-голямата част от биомедицинската ИИ като системи с висок риск, но неговата ефективност на практика ще зависи от това как гаранциите за основните права се прилагат при оценките на съответствието, мониторинга след пускането на пазара и обществените поръчки.

На първо място, гаранциите за значим човешки надзор трябва да бъдат засилени и конкретизирани за биомедицинските системи с ИИ през целия им жизнен цикъл. Клиничните инструменти с ИИ, използвани за диагностика, стратификация на риска, скрининг или подкрепа на лечението, в никакъв случай не трябва да функционират като де факто автономни вземащи решения. Човешкият надзор трябва да включва не само възможността за преодоляване от страна на медицинските специалисти, но и ясна институционална отговорност за разбиране на ограниченията на системата, известните рискове от предубеденост и разликите в ефективността на подгрупите. В съответствие с защитата на човешкото достойнство и неприкосновеност, заложена в Хартата, медицинските специалисти

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

трябва да бъдат обучени и да получават институционална подкрепа, за да подлагат на критичен анализ резултатите от изкуствения интелект, вместо да се подчиняват на тях. Това изисква включването на познания за изкуствения интелект, осведоменост за предубедеността и обучение по основните права в медицинското образование и продължаващото професионално развитие.

Задълженията за прозрачност следва да се тълкуват широко в контекста на здравеопазването. Пациентите и ползвателите на здравни услуги трябва да бъдат информирани винаги, когато системи с изкуствен интелект се използват при вземането на клинични решения, които ги засягат, включително при скрининг, определяне на приоритети или оценка на риска. Когато резултатите, генерирани от изкуствен интелект, се използват в обществените здравни услуги, тези резултати следва да бъдат ясно идентифицирани като такива и да бъдат придружени от достъпни обяснения за тяхната роля, ограничения и известни рискове от пристрастност. Лицата трябва да бъдат информирани и когато техните лични данни се използват за обучение, тестване или непрекъснато усъвършенстване на ИИ, особено когато става въпрос за чувствителни здравни данни. Тези мерки за прозрачност са от съществено значение за спазването на правата по Хартата за защита на данните и ефективни средства за защита, както и за да се даде възможност на лицата да оспорят по смислен начин решения, които могат да ги засегнат неблагоприятно.

Второ, оценката на въздействието върху основните права трябва да стане рутинно, приложимо изискване за биомедицинските системи за ИИ, като се разшири от проверки преди пускането на пазара до непрекъсната оценка по време на внедряването. Емпиричните данни в D2.1 показват, че много от вредите от предубедеността стават видими едва когато системите за ИИ взаимодействат с реални популации и клинични работни процеси, особено чрез пресечни ефекти, свързани с пола, расата, възрастта и социално-икономическия статус. Оценките на въздействието, основани на правата, като например тези, вдъхновени от методологията HUDERIA на Съвета на Европа (Методология за оценка на риска и

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

въздействието на системите с изкуствен интелект от гледна точка на правата на човека, демокрацията и върховенството на закона), следва да бъдат задължителни за медицинския ИИ с висок риск, като изрично се разглеждат разликите в представянето и резултатите между защитените групи. Тези оценки трябва да включват значимо участие на заинтересованите страни, включително организации на гражданското общество, представители на пациенти и органи за равенство, за да се открият вреди, които могат да бъдат невидими от чисто техническа или клинична гледна точка.

Следва да се изискват периодични одити на биомедицинските системи с изкуствен интелект, за да се провери продължаващото спазване на стандартите за основните права, като се обърне специално внимание на отклоненията в видовете предубеденост, промените в наборите от данни и промените в клиничното използване с течение на времето. Когато одитите разкрият трайни или неотстраними дискриминационни ефекти, трябва да има ясни правни и институционални пътища за ограничаване, суспендиране или прекратяване на използването на системата. Правото на здравеопазване не може да оправдае продължителното използване на инструменти с ИИ, които систематично поставят в неравностойно положение определени групи, дори ако общите показатели за ефективност изглеждат благоприятни.

Трето, европейските и националните органи трябва да се справят с риска от злоупотреба и вторични вреди, свързани с биомедицинския ИИ. Това включва уязвимости в киберсигурността, които могат да компрометират целостта на системата или да позволят злонамерена манипулация на клиничните резултати, както и пренасочването на ИИ в здравеопазването за наблюдение, профилиране или практики на изключване. Биомедицинските системи с ИИ трябва да подлежат на редовни оценки на сигурността и строги задължения за докладване на инциденти, с ясни механизми за отчетност в случаите, когато предубедени или компрометирани системи водят до нарушения на правата. Рамките за отговорност трябва да гарантират, че отговорността не може да се прехвърля единствено върху

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

отделни медици, когато вредите са структурно заложени в проектирането на ИИ или в решенията за внедряването ѝ.

Четвърто, насърчаването на етични и отговорни практики трябва да бъде заложено в цялата верига на стойността на биомедицинския ИИ. От разработчиците трябва да се изисква да се справят проактивно с рисковете от предубеденост чрез събиране на представителни данни, внимателен подбор на цели и прокси, валидиране за конкретни подгрупи и прозрачно отчитане на резултатите за различните полове и расови групи. Важно е, че прегледаните в D2.1 доказателства показват, че „равнопоставеността чрез неосведоменост“ и чисто техническите стратегии за премахване на предубедеността често са недостатъчни в здравните заведения. Следователно регулаторните насоки и стандарти трябва да надхвърлят абстрактните показатели за равнопоставеност и да изискват от разработчиците да демонстрират клинично значими резултати по отношение на равнопоставеността, оценени във връзка с реалните пътища на здравеопазването и моделите на достъп.

Политиките за обществени поръчки и финансиране играят ключова роля в оформянето на стимулите за разработчиците. Здравните власти и обществените болници трябва да интегрират основните права и критерии за непредубеденост в решенията си за възлагане на обществени поръчки за системи с ИИ, като дават предимство на решения, които демонстрират стабилни, прозрачни и независимо проверени практики за намаляване на предубедеността. Инструментите за финансиране на ЕС, включително бъдещите програми за научни изследвания и иновации, трябва да продължат да дават приоритет на проекти, които съчетават технически иновации с управление, основано на права, ангажираност на заинтересованите страни и изграждане на капацитет, в съответствие с модела AEQUITAS.

Накрая, засилването на обществената устойчивост към предубеден биомедицински ИИ изисква устойчиви инвестиции в осведомеността на обществеността, ангажираността на гражданското общество и междусекторното сътрудничество. Индивидите трябва да бъдат овластени да разберат своите права

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

в здравеопазването, подпомагано от ИИ, и механизмите, с които разполагат, за да ги защитят. Организациите на гражданското общество, органите за равенство и групите на пациенти трябва да бъдат признати като съществени участници в мониторинга на въздействието на ИИ, подкрепата на засегнатите лица и информирането за разработването на политики. Сътрудничеството между правителствата, доставчиците на здравни услуги, изследователите, индустрията и гражданското общество е необходимо, за да се гарантира, че ползите от биомедицинския ИИ се разпределят справедливо и не засилват съществуващите неравенства в здравеопазването.

Като цяло, заключенията от Доклад D2.1 подкрепят ясното политическо заключение: биомедицинският ИИ може да се счита за надежден и легитимен в ЕС само когато неговият дизайн, внедряване и управление се основават твърдо на защитата на основните права. Законът на ЕС за ИИ, тълкуван през призмата на Хартата на основните права на ЕС и приложен чрез конкретни механизми за надзор, оценка на въздействието и отчетност, предоставя критична възможност да се гарантира, че иновациите в здравеопазването насърчават равноправието, а не възпроизвеждат исторически модели на дискриминация.

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI

Съфинансирано от Европейския съюз

AEQUITAS

Финансирано от Европейския съюз. Изразените възгледи и мнения обаче принадлежат изцяло на техния(ите) автор(и) и не отразяват непременно възгледите и мненията на Европейския съюз или на Европейската изпълнителна агенция за образование и култура (EACEA). За тях не носи отговорност нито Европейският съюз, нито EACEA. Код на проекта: 101215009 — AEQUITAS — CERV-2024-CHAR-LITI