DOI 10.69085/PSC2025165

ДА РАЗБЕРЕШ НАБОРИТЕ ОТ ДАННИ: ПРЕДИЗВИКАТЕЛСТВА ПРИ ОБУЧЕНИЕТО НА ИЗКУСТВЕН ИНТЕЛЕКТ В ОБЛАСТТА НА СИГУРНОСТТА

проф. д-р Георги Димитров,
Университет по библиотекознание и информационни технологии, България

Деница Кожухарова, гл. юрисконсулт
Фондация „Право и Интернет“, България

Резюме: *Бързоразвиващите се технологии и новоприетата правна рамка на ниво Европейски съюз поставят редица предизвикателства пред внедрителите и доставчиците на системи, базирани на изкуствен интелект. Осигуряването на качествени набори от данни, с които се обучава изкуственият интелект, е от ключово значение за това той да отговори на изискванията за устойчивост, прозрачност и точност на системата. От изключително значение е да бъде проверен техният произход, начин на събиране или създаване, както и механизмът, по който се използват за обучаване на модела, базиран на изкуствен интелект – задължителни стъпки за гарантиране на етичното и отговарящо на правните норми изграждане на системата. В частност – в областта на сигурността, системи, базирани на изкуствен интелект се използват все по-широко от правоохранителните органи за оперативни цели и за целите на опазване на обществения ред. Част от тях биват класифицирани и като високорискови поради особеното засягане на неприкосновеността на личния живот на лицата, включително чрез профилиране и използване на големи информационни масиви (т. 6, Анекс III към Регламент (ЕС) 2024/1689). В този смисъл от особено значение е гарантирането на защитата на данните при проектирането на системата и задаването на подходящи правни изисквания към изграждането на софтуера.*

Ключови думи: *изкуствен интелект, набори от данни, сигурност*

DATASETS UNVEILED: CHALLENGES IN TRAINING AI FOR THE SECURITY DOMAIN

Prof. George Dimitrov
University of Library Studies and Information Technologies, Bulgaria

PhD, Denitsa Kozhuharova, legal advisor
Law and Internet Foundation, Bulgaria

Abstract: *The rapid evolution of technology, and the recently adopted legal framework at the European Union level, poses significant challenges for both deployers and providers of AI-based systems. Ensuring the quality of datasets used to train artificial intelligence is crucial for meeting the requirements of robustness, transparency, and accuracy. It is essential to verify the origin and method of collection or creation of these datasets, as well as the mechanisms though which they are used to train AI models. These steps are mandatory to ensure the ethical and legally compliant development of AI systems.*

In the security domain, AI-based systems are increasingly employed by law enforcement agencies for operational purposes and the maintenance of public order. These systems are often classified as high-risk due to their substantial impact on individual privacy, particularly through profiling and the use of big data (point 6, Annex III to Regulation (EU) 2024/1689). Therefore, it is imperative to ensure data protection by design and to establish appropriate legal requirements for the development of such software.

Key words: *artificial intelligence, datasets, security*

1. Увод

Развитието на изкуствения интелект (ИИ) през последните десетилетия доведе до значителни промени в различни сфери на обществения живот. В рамките на Европейския съюз (ЕС), създаването и приемането на нова нормативна уредба за регулиране на ИИ подчертава необходимостта от внимателно и отговорно внедряване на тези технологии. В този контекст настоящият доклад има за цел да анализира изискванията и необходимостта от наличието на гаранции, свързани с разработването и използването на системи, базирани на ИИ.

Ефективното и етично използване на ИИ изисква осигуряване на качествени набори от данни, които да служат за обучение на тези системи. Важността на данните и техният произход не могат да бъдат пренебрегнати, тъй като те играят ключова роля за постигане на прозрачност, точност и устойчивост на ИИ. Особено в сферата на сигурността и правоохранителната дейност, осигуряването на надеждни и качествени набори от данни е от критично значение.

Настоящият доклад разглежда различните видове набори от данни, които са необходими за жизнения цикъл на ИИ системи и подчертава правните и етични аспекти, свързани с тяхното използване. Анализът се съсредоточава върху особеностите на новоприетото европейско законодателство и предлага практически насоки за разработчиците и доставчиците на системи с ИИ. Докладът изследва още ролята на ИИ в областта на сигурността, като поставя акцент върху изискванията, приложими към високорисковите системи и необходимостта от защита на личните данни.

Създаването на този доклад е финансирано от Европейския съюз. Изразените възгледи и мнения принадлежат единствено на авторите и не отразяват непременно тези на Европейския съюз или Европейската комисия. Нито Европейският съюз, нито Европейската комисия носят отговорност за тях.

2. Изясняване на ключови понятия

С оглед на провеждането на кратко обследване на предизвикателствата, породени от разработването на системи с изкуствен интелект (ИИ) в сферата на сигурността и борбата с престъпността, на първо място е редно да бъдат изяснени някои ключови понятия. По-долу са разгледани двата основни термина, които следва да бъдат добре разбрани, така че настоящият доклад да бъде осмислен в своята цялост.

2.1. Изкуствен интелект

Понятието „изкуствен интелект" наскоро получи легална дефиниция в рамките на Акта на ЕС за ИИ (Регламент 2024/1689), която определя, че система с ИИ е „*машинно базирана система, която е проектирана да работи с различни нива на автономност и която може да прояви адаптивност след внедряването си и която, с явна или с подразбираща се цел, въз основа на въведените в нея входящи данни, извежда начина на генериране на резултати като прогнози, съдържание, препоръки или решения, които могат да окажат влияние върху физическа или виртуална среда*".[1] Въпреки това както текстът на цитираната разпоредба, така и официалният му превод на български език може да остави известни съмнения или пък да постави трудности при пълното разбиране на термина. В тази връзка следва да се има предвид следното определение, предоставено в Етични насоки за надежден изкуствен интелект от експертната група на високо равнище по въпросите за изкуствения интелект, а именно *„[с]истемите за ИИ са софтуерни (а може би и хардуерни) системи, създадени от хора, които при зададена сложна цел действат във физическото или цифровото измерение като разпознават средата си чрез събиране на данни, тълкувайки събраните структурирани или неструктурирани данни, въз основа на знание или обработване на информация, получена от тези данни, като вземат най-доброто решение, което да се предприеме за постигане на дадената цел. Системите с ИИ могат да използват знакови правила или да научат цифров модел, а*

[1] Вж. чл. 3, § 1, Регламент (ЕС) 2024/1689 на Европейския парламент и на Съвета от 13 юни 2024 година за установяване на хармонизирани правила относно изкуствения интелект (Акт за изкуствения интелект), Официален вестник на ЕС, серия L от 12 юли 2024 г.

също така могат и да адаптират поведението си като анализират как средата е повлияна от предишните им решения".[2]

Видно от цитираните определения, системите с ИИ се характеризират със своята самостоятелност и независимост при работата си, като начина за вземане на решения следва да бъде максимално разбираем и прозрачен за хората, които работят с тези системи. На практика това означава хората не просто да следят процесите, които се случват в рамките на дадена система с ИИ, но и да умеят да разбират как се задействат тези процеси, какви фактори се вземат предвид при формирането на решения и предложения, както и какви други мотиви има дадената система с ИИ при формулирането на решения. Дефинициите същевременно сочат и важната роля, която играят данните, които се обработват за решаване на поставените задачи пред системите с ИИ. В следващата подточка са разгледани отделните категории от данни, които са необходими за жизнения цикъл на всяка система с ИИ.

2.2. Видове и набори от данни

Макар и определенията по-горе да говорят за данни, на практика за създаването, обучението и потвърждението на всяка една система с ИИ се изисква работа с т. нар. набори от данни. Наборът от данни е структурирана съвкупност от данни, обикновено организирана по начин, който я прави лесно достъпна, управляема и анализируема.[3]

С оглед създаването на която и да е система с ИИ, е необходимо процесът по нейното разработване да стъпи върху различни видове набори от данни, а именно:

- **обучителни данни**, които се използват за същинското обучение на система с ИИ[4]. Именно това е и най-важната категория данни, тъй като на нейната база се оформя функционирането на дадена система с ИИ по отношение на вземането на решения. При пропуски в тези набори от данни, непълнота, грешки или в неблагоприятен случай – въвеждане на данни, възпроизвеждащи предразсъдъци или дискриминационни възгледи, това може да доведе и до неправилно и неетично вземане на решения. Последното би могло да има сериозни последици върху различни житейски сфери на засегнатите физически лице (пр. отказ на достъп до социални и здравни услуги, неправилно оценяване, криминално профилиране и др. под.);

[2] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, 2019, p. 36.

[3] Snijders, C. et al. 'Big Data': Big gaps of knowledge in the field of Internet. – International Journal of Internet Science, 2012, No 7, 1–5.

[4] Чл. 3, § 29, Регламент (ЕС) 2024/1689.

• **валидационни данни**, които служат за оценка на обучената система с ИИ и за допълнителна настройка на някои нейни параметри, които не подлежат на машинно обучение. Те служат още за оценка на самия обучителен процес с цел наред с другото да се предотврати недостатъчно или прекомерно настройване на системата с ИИ.[5] За целта може да е необходимо отделянето/създаването на самостоятелен **набор от валидационни данни**[6]. Подобно на обучителните данни, качеството, разнообразието, представителността и точността на наборите от валидационните данни са с голямо значение за законосъобразното и етично функциониране на системите с ИИ. Добра практика е наборите от валидационни данни да са с различен произход и източник от наборите от обучителни данни, за да се постигне добро сравнение и реално удостоверяване, че системата с ИИ работи по планирания начин и произвежда желани, но и коректни резултати.

• **изпитвателни данни** се използват за извършване на независима оценка на системата с ИИ с цел потвърждаване на очакваното действие на тази система преди пускането ѝ на пазара/в действие.[7] Както е посочено, тези набори от данни се въвеждат в дадена система с ИИ непосредствено преди да започне нейното оперативно и ежедневно използване. Ролята на изпитвателните данни е да дадат гаранции, че системата работи законосъобразно и етично спрямо проектираните функционалности. Изпитвателните данни могат напълно да „имитират" входящи данни, за да симулират създаването на различни задачи за решаване от дадената система с ИИ, за да се постигне извършването на тази независима оценка.

• **входящи данни** са онези данни, които се предоставят на система с ИИ или се придобиват пряко от нея, въз основа на които системата произвежда резултат.[8] Това са и данните, визирани във второто определение, посочено в предходната подточка. Именно това са данните, с които пряко работи дадена система с ИИ. Те биват обработвани, оценявани, разглеждани в контекста на конкретната задача, която система с ИИ решава, както и в корелация с други фактори, които да насочат към вземането на решение или пък предоставянето на препоръка. Входящите данни могат да бъдат въведени пряко от лицата, използващи системата с ИИ, но и могат да бъдат събрани/изискани по автоматизиран път (пр. при използване на система с

[5] чл. 3, пар. 30, Регламент (ЕС) 2024/1689.
[6] чл. 3, пар. 31, Регламент (ЕС) 2024/1689.
[7] чл. 3, пар. 32, Регламент (ЕС) 2024/1689.
[8] чл. 3, пар. 33, Регламент (ЕС) 2024/1689.

ИИ в областта на наказателния процес системата може да извлече необходимите данни от релевантните регистри, поддържани от министерството на вътрешните работи или от министерство на правосъдието).

В рамките на отделните набори от данни е възможно да присъстват и данни от личен характер. Те биват:

- „обикновени" лични данни – всяка информация, която може да доведе до идентифицирането на дадено физическо лице, пряко или непряко, по-специално чрез идентификатор като име, идентификационен номер, данни за местонахождение, онлайн идентификатор.[9]
- специална категория лични данни („чувствителни" данни) – информация, разкриваща расов или етнически произход, политически възгледи, религиозни или философски убеждения или членство в синдикални организации, както и генетични данни, биометрични данни за целите единствено на идентифицирането на физическо лице, данни за здравословното състояние или данни за сексуалния живот или сексуалната ориентация на физическото лице.[10] На практика това са тези категории от данни, чието незаконосъобразно обработване би могло да доведе до възникване на дискриминация.

В контекста на разработването на системи с ИИ в сферата на сигурността и борбата с престъпността качеството и точността на наборите от данни са от изключително значение, тъй като системите могат да бъдат проектирани, разработени и изпитани по начин, който гарантира пълното зачитане на правата на лицата, щом дадена система с ИИ започне да се прилага в действие.[11]

3. Нормативна уредба на изкуствения интелект. Приложими етични правила

3.1. Регламент (ЕС) 2024/1689

Пътят към приемането на Регламент (ЕС) 2024/1689 на Европейския парламент и на Съвета от 13 юни 2024 година за установяване на хармонизирани правила относно изкуствения интелект и за изменение на

[9] чл. 4, пар. 1, Регламент (ЕС) 2016/679 на Европейския парламент и на Съвета от 27 април 2016 година относно защитата на физическите лица във връзка с обработването на лични данни и относно свободното движение на такива данни (Общ регламент относно защитата на данните), Официален вестник на ЕС, серия L 119/1 от 4 май 2024 г.

[10] чл. 9, пар. 1, Регламент (ЕС) 2016/679.

[11] Kusak, M. Quality of data sets that feed AI and big data applications for law enforcement. – ERA Forum 23, no. 2 (October 1, 2022): 209–19, p. 210.

регламенти (ЕО) № 300/2008, (ЕС) № 167/2013, (ЕС) № 168/2013, (ЕС) 2018/858, (ЕС) 2018/1139 и (ЕС) 2019/2144 и директиви 2014/90/ЕС, (ЕС) 2016/797 и (ЕС) 2020/1828 (Акт за изкуствения интелект) започва на 21 април 2021 г. Периодът за обществени консултации приключва през август същата година, когато е и публикувано проучване от етична и правна гледна точка относно обработването на биометрични данни.[12] През декември 2022 г. Съветът на ЕС приема общата си позиция по Акта за ИИ като предлага хармонизиране на националните правила за отговорност за ИИ, а няколко месеца по-рано – през септември 2022 г., Комисията по правни въпроси в Европейския парламент приема становище относно Акта за ИИ. Това води до гласуване на проекта за регламент в Европейския парламент през юни 2023 г., приет с 499 гласа „за" и гласуване от 27-те държави членки на ЕС, които единодушно одобряват Акта за ИИ през декември 2023 г. Актът за изкуствения интелект на Европейския съюз е официално приет през 2024 г., а Европейската служба за ИИ стартира през февруари 2024 г., за да подпомогне практическото приложение на Акта.[13]

Една от отличителните черти на Акта за ИИ е възприетият подход, основан на риска. Спрямо него разпоредбите на регламента се прилагат в зависимост от установеното ниво на риск. Системите за ИИ, които биха могли да имат **значително отрицателно въздействие върху здравето, безопасността или правата на човека**, се класифицират като **високорискови** и изискват покриването на редица условия, преди да могат да бъдат предложени на единния цифров пазар на ЕС. Системите за ИИ, които страдат от **липса на прозрачност**, се класифицират като **среднорискови**. Подобно на високорисковите системи за ИИ, към тях също е предвиден определен набор от задължения, макар и с по-нисък интензитет. **Системите за ИИ, които не се класифицират нито като високорискови, нито като среднорискови**, се считат за носещи **минимален риск** и не подлежат на допълнителни условия.

Високорискови системи с ИИ

Приема се, че високорисковите системи с ИИ могат да представляват значителен риск за здравето, безопасността или основните права на гражданите на ЕС, но чиито социално-икономически ползи

[12] Wendehorst C. et al., Biometric Recognition and Behavioural Detection. Policy Department for Citizens' Rights and Constitutional Affairs, 2021 (https://www.europarl.europa.eu/RegData/etudes/STUD/2021/696968/IPOL_STU(2021)696968_EN.pdf).

[13] Historic Timeline | EU Artificial Intelligence Act, a.n.d. (https://artificialintelligenceact.eu/developments/).

превъзхождат тези рискове.[14] По-долу изчерпателно са посочени всички видове високорискови системи с ИИ:

- Система с ИИ, която обработва **биометрични данни** и извършва: дистанционна биометрична идентификация (с изключение на елементарни системи за биометрична проверка с ИИ, например системите за граничен контрол, внедрени на летища), която не се извършва в реално време; категоризация въз основа на чувствителни данни; емоционално разпознаване.[15]
- Системи с ИИ, използвани в областта на сигурността на **критична инфраструктура** от гледна точка на управление и експлоатация.[16]
- Системи с ИИ в областта на **образованието и професионалното обучение**, използвани за: прием; оценяване; определяне на ниво на образование; установяване на неразрешено поведение на ученици/студенти по време на изпит.[17]
- Системи с ИИ в областта на **заетостта**, използвани за: улесняване на процесите по набиране на персонал чрез публикуване на обяви за работа, филтриране и анализ на кандидатурите и оценка на кандидатите; вземане на решения на работното място, като например повишаване и понижаване в длъжност, прекратяване на трудов договор, оценка на изпълнението на служебни задължения въз основа на индивидуални характеристики и лични особености.[18]
- Системи с ИИ в областта на **услугите**, които се използват за: оценка за достъп до услуги и помощи, предоставяни от публични органи (пр. социални и здравни услуги и др. под.); определяне на кредитен рейтинг; определяне на цените на здравно застраховане и животозастраховане; триаж на спешни случаи.[19]
- Системи с ИИ, които се използват в **правоохранителната сфера** и които се използват от самите правоохранителни органи или от тяхно име и чието използване е предвидено в приложимото право за целите на: оценка на риска дадено лице да стане жертва на престъпление; разпознаване на лъжа и други подобни човешки поведения; оценка на доказателствата; оценка на вероятността от повторно из-

[14] Gehrmann, M. C., A. Förster A. AI Act: high-risk AI systems – what applies, what is due when? – *Taylor Wessing*, 2024 (https://www.taylorwessing.com/en/insights-and-events/insights/2024/11/high-risk-ai-systems).
[15] Прил. III, § 1, Регламент (ЕС) 2024/1689.
[16] Прил. III, § 2, Регламент (ЕС) 2024/1689.
[17] Прил. III, § 3, Регламент (ЕС) 2024/1689.
[18] Прил. III, § 4, Регламент (ЕС) 2024/1689.
[19] Прил. III, § 5, Регламент (ЕС) 2024/1689.

*върш*ване на престъпление; профилиране в областта на наказателните дела.[20]

- Системи за ИИ, използвани в областта на **граничния контрол, убежището и миграцията**, съгласно приложимото право с оглед на: разпознаване на лъжа и други подобни човешки поведения от публичните органи; оценка на здравен риск от компетентните органи; оценка на риска за сигурността (вкл. риска от неправомерна миграция) от компетентните органи; оценка на молбата за получаване виза/убежище/постоянно пребиваване (вкл. предоставената съпътстваща документация) от компетентните органи; откриване/разпознаване/идентифициране на физически лица от компетентните органи. Това не включва обикновена проверка на документи за пътуване.[21]
- Системи с ИИ в областта на **правосъдието** и **изборите**, които: подпомагат съдебните органи при решаване на спорове (включително алтернативни методи за решаване на спорове) по отношение на тълкуване на факти, юридически проучвания, прилагане на закона към съответния случай; влияят на индивидуалното поведение при гласуване на дадено лице или на цялостния резултат от съответните избори/референдум. Това не включва използването на ИИ в административната и материално-техническа подготовка на изборите.[22]

С оглед специфичната тема на настоящия доклад, интерес представляват системите с ИИ, приложими в правоохранителната сфера и тези в областта на граничния контрол. Видно от изложената информация, има голяма вероятност подобни системи да бъдат класифицирани като високорискови, тъй като при неточни резултати или пък такива, възпроизвеждащи предразсъдъци, може съществено да бъде засегната личната сфера на лицата в значителен размер (пр. дадено лице да бъде несправедливо обвинено в извършване на престъпление или да се направи погрешна категоризация и оценка на доказателства, която да оневини виновно лице). Следващите точки и подточки на доклада ще разгледат специфичните изисквания към подобни системи с ИИ именно в тази хипотеза.

Нормативни изисквания спрямо наборите от данни

В рамките на Акта за изкуствен интелект (чл. 10) са предвидени нарочни разпоредби, които уреждат изискванията спрямо наборите от данни, използвани в контекста на високорискови системи с ИИ. На първо място, наборите от обучителни, валидационни и изпитателни данни

[20] Прил. III, § 6, Регламент (ЕС) 2024/1689.
[21] Прил. III, § 7, Регламент (ЕС) 2024/1689.
[22] Прил. III, § 8, Регламент (ЕС) 2024/1689.

следва да са подходящи, достатъчно представителни, във възможно най-голяма степен без грешки и пълни с оглед на специфичното предназначение на системата. Необходимо е наборите от данни да притежават целесъобразни статистически характеристики, включително, когато е уместно, спрямо индивидите или групите, за които високорисковата система с ИИ ще бъде приложима. Препоръчително е тези характеристики на наборите от данни да са налице както с оглед на отделните набори от данни, така и по отношение на комбинация от тях[23]. В зависимост от необходимостта, свързана с предназначението на дадената високо рискова система с ИИ, наборите от данни трябва да бъдат адаптирани спрямо особеностите или елементите, присъщи на конкретната среда, в която тя ще се приложи от гледна точка на география, поведения, функционалност и контекст.[24]

Налице са някои норми, които регулират обработването на лични данни от високорискови системи с ИИ. Така например Актът за ИИ гласи, че по изключение могат да бъдат обработвани чувствителни лични данни при установяването на подходящи гаранции за основните права и свободи на физическите лица. По-долу са посочени кумулативно установените от законодателя изисквания за обработване на специални категории лични данни:

- налице е правно основание по смисъла на Общия регламент за защита на данните или Директива (ЕС) 2016/680 на Европейския парламент и на Съвета от 27 април 2016 година относно защитата на физическите лица във връзка с обработването на лични данни от компетентните органи за целите на предотвратяването, разследването, разкриването или наказателното преследване на престъпления или изпълнението на наказания и относно свободното движение на такива данни;
- откриването и коригирането на предразсъдъци не може да се осъществи ефективно чрез обработването на други данни, включително на такива, които са създадени по синтетичен път или на т.нар. анонимизирани данни;
- прилагат се технически ограничения, свързани с повторното използване на лични данни, както и най-съвременни мерки за сигурност и опазване на неприкосновеността на личния живот, включително псевдонимизация;
- прилагат се мерки, с които се гарантира, че обработваните лични данни са обезопасени, защитени, спрямо тях се прилагат подходящи гаранции, включително строг контрол и документиране на достъпа.

[23] Чл. 10, § 3, Регламент (ЕС) 2024/1689.
[24] Чл. 10, § 4, Регламент (ЕС) 2024/1689.

Такива мерки могат да бъдат например: ясни вътрешни правила относно събирането и обработването на лични данни, приложими към служителите, отговорни за разработването, обучението, валидирането и изпитването на система с ИИ, завишен контрол на достъп до тези данни, ограничен както по отношение на времето, така и по отношение на лицата, провеждането на начални и последващи обучения относно етичните аспекти на ИИ с оглед изостряне на вниманието на разработчиците и техните ръководители относно потенциалните неблагоприятни последици, които могат да възникнат при неправилни резултати от работата на системата с ИИ, воденето на подробни регистри относно процесите по обработване на данни, но и по повод на обмена на данни между различните компоненти на системата с ИИ, писмено документиране на процеса по изпитване, регулярно одитиране на процеса по разработване, но и на самата система с ИИ по установена методология от независима организация/лица и др. под.

- специалните категории лични данни не се предоставят, не се предават, нито се достъпват по друг начин от други лица, на които подобен достъп не се предоставя;
- специалните категории лични данни се заличават, след като предубежденията на системата с ИИ са коригирани или е достигнат краят на срока на съхранение на личните данни. Препоръчително е този процес да бъде стриктно документиран;
- документирано е защо обработването на специални категории лични данни е строго необходимо за откриване и коригиране на предубеденост на ИИ и не било възможно тази цел да се постигне чрез обработване на други видове (лични) данни.[25]

Всички тези препоръки следва да бъдат претворени спрямо процеса на проектиране, изпитване и експлоатиране на конкретната система с ИИ в подходящи вътрешни правила и процедури, така че да се осигури надлежното документиране на всяка една от посочените по-горе стъпки. Изчерпателно следва да се посочат всички технологии, методики и сертификационни схеми, които ще се използват за постигане на съответствие с посочените условия за обработване на специални категории лични данни. При предлагането на пазара на система с ИИ, която като част от входящите данни ще обработва специална категория данни, следва препоръките и инструментите, в смисъла на постигане на такова документиране. Тези препоръки и инструменти са част от услугата, предлагана на клиента.

[25] Чл. 10, § 5, Регламент (ЕС) 2024/1689.

3.2. Етична рамка на Европейския съюз относно Изкуствения интелект – Етични насоки за надежден изкуствен интелект, предоставени от експертната група на високо равнище по въпросите за изкуствения интелект

Етичните насоки за надежден ИИ, публикувани през 2019 г. от Експертната група на високо равнище по изкуствен интелект, очертават рамка, чието практическо приложение гарантира, че системите за ИИ се разработват и използват по законосъобразен, етичен и устойчив начин. Насоките подчертават значението на ИИ, който поставя човека в центъра на вземането на решения, насърчавайки човешкото благосъстояние и основните права. Насоките установяват ключови изисквания, чието наличие прави дадена система с ИИ етична, включително от гледна точка на принципите за прозрачност, отчетност и справедливост. Те подчертават и необходимостта от непрекъснато наблюдение и оценка на системите с ИИ, за да се гарантира зачитането на тези принципи. Насоките обсъждат и значението на техническата надеждност и безопасност на системите с ИИ. Това включва осигуряване на сигурност, надеждност и устойчивост на системите спрямо атаки и грешки. Освен това насоките призовават за въвеждането на механизми, които да позволяват реален човешки надзор и намеса, осигурявайки възможност за контрол и корекция на системите с ИИ, когато е необходимо. Това на практика означава всички крайни решения да бъдат взети от човек, а не от машина, тъй като не е достатъчно човекът да бъде единствено информиран за работата и логиката на произвеждането на резултати от дадена система с ИИ. Необходимо е човекът да бъде овластен да взема решения, макар и на база на препоръки в резултат от анализа на дадена система с ИИ. Това е и златното правило, чието спазване се насърчава в сферата на сигурността и борбата с престъпността и тероризма.

Що се касае до специфичната тема на доклада, насоките препоръчват използването на висококачествени набори от данни и провеждането на взискателни изпитвания на системите с ИИ, за да се намалят максимално всякакви пристрастия и неточности. В тази връзка може да се мисли в посока на обособяване на екипи вътре в самата наказателноправна/правоохранителна система относно създаване на качествени набори от данни и дори за проектиране, изпитване и оперативно прилагане на системи с ИИ. По този начин ще се осигури разнообразие от набори данни, които по законосъобразен ред могат да бъдат събирани и съхранявани единствено от компетентните органи. Като първа стъпка може да се насърчи по-всеобхватното събиране на статистически данни от институциите, които имат вменено подобно задължение. Освен за

изготвянето на висококачествени и представителни набори от данни, това ще позволи и по-доброто вземане на управленски решения.

Не на последно място, насоките разглеждат социалното и екологичното въздействие на ИИ. Те насърчават разработването на системи с ИИ, които допринасят положително за обществото и околната среда, насърчавайки устойчивост и социално включване. Насоките също така подчертават необходимостта от сътрудничество между различни заинтересовани страни, включително законодатели, разработчици на такива системи и потребители, за създаване на всеобхватен и приобщаващ подход за управление на ИИ. Включването на такъв широк спектър от граждани е мислимо в контекста на изготвянето на различни методики за одитиране и сертифициране на системи с ИИ, така че гледната точка на лицата, които могат да бъдат засегнати от резултатите на дадена система с ИИ, да стане неразривна част от процеса по оценяването ѝ или пък да спомогне за изграждането на критично мислене у самите разработчици и ползватели. За авторите на насоките това съвместно усилие се счита за съществено за изграждането на обществено доверие и гарантиране, че системите с ИИ са в съответствие с обществените ценности и етични стандарти.

В насоките са установени три основни елемента, правещи всеки ИИ надежден:

- законосъобразност, което означава, че ИИ трябва да е в съответствие с приложимото законодателство,
- етичност, което означава, че ИИ трябва да се придържа към етичните ценности и принципи, и
- надеждност, което означава, че ИИ трябва да бъде разработен по начин, който да сведе до минимум потенциалните вреди от техническа и социална гледна точка.[26]

В насоките се посочват седем ключови изисквания за надежден ИИ: човешко участие и надзор, техническа надеждност и безопасност, неприкосновеност на личния живот и управление на данните, прозрачност, многообразие, недискриминация и справедливост, обществено и екологично благополучие и отчетност. Тези изисквания имат за цел да гарантират, че системите за ИИ овластяват хората, че са сигурни и надеждни, зачитат неприкосновеността на личния живот, че са прозрачни в своите операции, насърчават справедливостта и приобщаването, допринасят положително за обществото и околната среда и носят отговорност за своите действия и решения.[27]

[26] High-Level Expert Group on Artificial Intelligence, Op. cit., p. 5.
[27] High-Level Expert Group on Artificial Intelligence, Op. cit., p. 14.

Тъй като тези насоки времево изпреварват законодателните усилия на органите на Европейския съюз, дълго време на практика те бяха единствената отправна точка, ръководеща създаването, изпитването и практическото приложение на ИИ в сферата на сигурността. Чисто концептуално, много от препоръките, посочени в насоките, са възприети в Акта за ИИ. Въпреки това, когато се касае до високорискови системи с ИИ в областта на сигурността, освен спазването на нормативните изисквания, силно препоръчително е и следването на насоките, за да се гарантират в максимален обем правата на лицата.

4. Ролята на наборите от данни при обучение и утвърждение на системи с изкуствен интелект. Спецификите в сферата на сигурността

Както бе посочено по-горе, почти всяка система с ИИ, която се прилага в сферата на борбата с престъпността и обществената сигурност, вероятно ще бъде класифицирана като високорискова. За тази цел отговорното разработване на подобни технологии, според новото законодателство, включва стриктно управление на риска[28], специално тестване и обучение, процедури за управление на данните[29], спазване на стандартите за киберсигурност[30] и осигуряване човешки надзор[31].

Каква е ролята на наборите от данни за спазването на тези принципи? На първо място, както бе обсъдено и по-горе, качеството на наборите от данни е от решаващо значение, тъй като това определя и качеството на резултатите, които могат да се постигнат от системата с ИИ[32]. Използването на ИИ при борбата с престъпността и тероризма е свързано с високи рискове, ако за обучение, валидация и изпитване са използвани непълни или некачествени набори от данни. Един от основните рискове е получаването на предубедени резултати или пък възникването на дискриминация[33].В наказателно-правен контекст това може да доведе до неправилното установяване на лице или група от лица, които да представляват интерес и обект за обследване от органите на реда, несправедливо обвинение или неточности в резултатите

[28] Чл. 9, Регламент (ЕС) 2024/1689.
[29] Чл. 10, Регламент (ЕС) 2024/1689.
[30] Чл. 15, Регламент (ЕС) 2024/1689.
[31] Чл. 14, Регламент (ЕС) 2024/1689.
[32] FRA. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. – FRA Focus, 2021, p. 3 (https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect).
[33] FRA, Op. cit., p. 8.

от криминалистичната експертиза, отпадане на качеството „пострадал" спрямо определено лице. Използване на подобни алгоритми би могло да накърни още и правото на свободно изразяване[34], ако в резултатите има грешки или пък са разтълкувани неправилно, когато ИИ се използва за идентифициране на терористично съдържание онлайн, радикализация в интернет или за подбуждане към престъпление.

За да се избегнат рискове за правата и свободите на гражданите, създателите и потребителите на подобни инструменти с ИИ следва да въведат подходящи и пропорционални мерки и гаранции, като например:

- **надеждни набори от данни** както за разработването, така и за изпитването на инструментите, като се уверят, че избраните набори от данни не възпроизвеждат пристрастия и не съдържат дискриминационни наклонности.[35] Тази насока следва да се прилага както спрямо наборите от реални данни, така и спрямо набори от синтетични данни. Наборите от данни следва да са максимално подробни, съдържащи голям брой артефакти и детайли от обективната реалност, които имат практическо решение за проектираната система с ИИ. Те следва да имат подходящ обхват (географски, тематичен, езиков и др.) Така например при разработване на системи за разпознаване на терористично съдържание, които ще се прилагат на територията на ЕС, не е редно да се вземат предвид набори от данни, чието съдържание се отнася до Северна Америка или Югоизточна Азия или пък такива, чието съдържание е на езици, които не се използват на територията на ЕС.
- участие на **разнообразен екип** в проектирането/разработването на съответната система за ИИ.[36] Тази препоръка е от особено значение за избягването на риска от резултати, които съдържат предубеждения или водят до дискриминация. Тя следва да се разбира като създаването на екип, който включва представители на различни полове, вероизповедания, етнически групи и националности.
- гарантиране, че **хората, които използват системата на ИИ,** не само познават методиката на вземане на решения, но и активно участват във вземането им.[37] Тази препоръка е в синхрон с изразеното по-горе становище, че е необходимо човекът да е овластен да

[34] Dietrich, F. AI-based removal of hate speech from digital social networks: chances and risks for freedom of expression. - AI And Ethics, 2024, 4–5 (https://link.springer.com/article/10.1007/s43681-024-00610-7)
[35] FRA, Op. cit., p. 13.
[36] High-Level Expert Group on Artificial Intelligence, Op. cit., p. 19.
[37] High-Level Expert Group on Artificial Intelligence, Op. cit., p. 12.

взема решения въз основа на аналитичната работа на системата с ИИ, а не сляпо да се доверява на постигнатите от машината резултати.

- **надлежно документиране** на всяка дейност, извършвана както от човека, така и от машината, за по-добра прозрачност.[38] Макар това да води до увеличаване на административната тежест, тази препоръка е важна за пълната проследяемост на процесите относно вземането на решение или формулиране на препоръки. Мислимо е да се извърши автоматизиране (пр. водене на регистри) на този процес с оглед неговото улесняване.
- осигуряване на **първоначално и продължаващо обучение** по отношение на това как трябва да се тълкуват резултатите, получени от съответната система с ИИ, и какви са нейните ограничения[39]. Както е посочено по-горе, развиването на критично мислене у лицата, които разработват и прилагат система с ИИ, е от ключово значение за нейното етично използване. Важно е не само да се постигне разбиране на процесите, които протичат при формулирането на резултат, но и познаването на това какви са първоначалните цели, заложени на етап проектиране на дадената система, какви набори от данни са използвани за обучение, валидиране и изпитване, какви са подходящите вътрешни правила и мерки за етичното боравене с дадената система и др. под.
- развитие на **критично мислене** по отношение на неблагоприятните последици, които могат да настъпят в случай на вземане на погрешно решение.[40] Тази препоръка върви ръка за ръка с предходната подточка, тъй като визира необходимостта от активно участие на човека при вземането на решение с подкрепата на анализи от ИИ, а не сляпото (пре)доверяване на предложенията и препоръките, които машината е способна да направи. Това важи с още по-голяма сила за системите с ИИ, които се използват в контекста на обществената сигурност и борбата с престъпността, тъй като последиците от неправилното, незаконосъобразното и неетично прилагане на систе-

[38] Захариев М. Правни гаранции за сигурността на личните данни, обработвани от компетентните органи за полицейски и наказателни дейности. – В: Сигурност и отбрана, С., АИ За буквите (Zahariev M. Pravni garantsii za sigurnostta na lichnite danni, obrabotva-ni ot kompetentnite organi za politseyski i nakazatelni deynosti. – V: Sigurnost i otbrana, S., AI Za bukvite), 2023, 492–506.

[39] Murire, O. T. „Artificial Intelligence and Its Role in Shaping Organizational Work Practices and Culture." Administrative Sciences 14, 2024, no. 12, p. 10 (https://www.mdpi.com/2076-3387/14/12/316).

[40] High-Level Expert Group on Artificial Intelligence, Op. cit., p. 13.

ма с ИИ могат да доведат до грубо вмешателство в най-фундаменталните права на лицата.

5. Заключение

Макар на пръв поглед появата на системи с ИИ да изглежда като панацея за редица проблеми, свързани с осигуряването на обществената сигурност и борбата с престъпността и тероризма, анализът в настоящия доклад ясно показва, че е необходимо преди всичко да се мисли в посока как човекът, оперирайщ машината, следва да тълкува и прилага получените резултати. Приемането на нормативна уредба от страна на органите на ЕС е важна крачка за правилното функциониране и прилагане на системи с ИИ, тъй като създава сигурност у гражданите, че предвидените гаранции относно правата и свободите им, които им се гарантират като жители на ЕС, ще бъдат зачетени на всяка стъпка от жизнения цикъл на дадена система с ИИ.

Един от фундаментите на подобна гаранция е именно правилният подбор на набори от данни, необходими за обучение, валидиране, изпитване и същинска работа на дадена система с ИИ. Наличието на разнообразни, представителни, качествени, точни и прецизни набори от данни е една от основните гаранции за правата и свободите на гражданите на ЕС. Това важи с още по-голяма сила, когато система с ИИ се прилага за опазване на обществения ред, борбата с престъпността и тероризма.

Въпреки че Актът за ИИ е основният законодателен инструмент, който ще навигира проектирането, разработването и прилагането на ИИ в рамките на ЕС, съществуващите етични насоки не следва да бъдат пренебрегнати. Напротив, те дават ценна насока относно това какви са ценностите, насърчавани от ЕС, които следва да бъдат претворени в технически изисквания спрямо системите с ИИ, използвани на територията на Съюза.

Системите с ИИ са мощен инструмент за решаването на задачи и отключването на анализаторски потенциал, особено в сферата на сигурността и правоохранителната дейност. Не бива да се забравя обаче, че отговорното и етично приложение на подобен инструмент все пак изисква активно поведение от страна на човека – неговото овластяване, насърчаване на критичното мислене и продължаващо обучение.