ЕКСПЕРТНА ГРУПА НА ВИСОКО РАВНИЩЕ ПО ВЪПРОСИТЕ НА ИЗКУСТВЕНИЯ ИНТЕЛЕКТ

към Европейската комисия

ПРОЕКТ НА
НАСОКИ ОТНОСНО ЕТИЧНИТЕ АСПЕКТИ ЗА НАДЕЖДЕН ИИ

РЕЗЮМЕ

Работен документ за консултация със заинтересованите страни

Брюксел, 18 декември 2018 г.

ПРОЕКТ НА НАСОКИ ОТНОСНО ЕТИЧНИТЕ АСПЕКТИ

ЗА НАДЕЖДЕН ИИ

Експертна група на високо равнище по въпросите на изкуствения интелект
Проект на насоки относно етичните аспекти за надежден ИИ

Европейска комисия
Генерална дирекция „Комуникации“

За контакти Nathalie Smuha — координатор на експертната група на високо равнище по въпросите на изкуствения интелект
Електронна поща: CNECT-HLG-AI@ec.europa.eu

Европейска комисия
B-1049 Брюксел

Документът е представен публично на 18 декември 2018 г. на английски език.

Този работен документ бе изготвен от експертната група на високо равнище по въпросите на изкуствения интелект, без да се засягат индивидуалните позиции на нейните членове по конкретни въпроси и без да се засяга окончателната редакция на документа. Той ще бъде доразработен, като окончателната му редакция ще бъде представена през март 2019 г. след консултацията със заинтересованите страни чрез Европейския алианс за ИИ.

Повече информация за експертната група на високо равнище по въпросите на изкуствения интелект ще намерите на следния адрес: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

РЕЗЮМЕ

Настоящият работен документ представлява проект на насоки относно етичните аспекти във връзка с ИИ, изготвени от експертната група на високо равнище по въпросите на изкуствения интелект (ЕГВР ИИ) към Европейската комисия. Окончателната редакция на документа трябва да бъде изготвена през март 2019 г.

Изкуственият интелект (ИИ) е един от факторите, водещи до най-големи трансформации в днешно време, и неминуемо ще доведе до преобразяване на обществото. Той разкрива отлична възможност за постигане на повече благополучие и по-висок растеж, което е неминуемо една от целите на Европа. Благодарение на огромните количества цифрови данни, мощните изчислителни архитектури и развитието в техниките на основата на ИИ, като например машинното самообучение, през последното десетилетие бе постигнат сериозен напредък в редица области. С всеки изминал ден основаващи се на ИИ сериозни разработки в областта на автономните превозни средства, здравеопазването, домашните и служебните роботи, образованието и киберсигурността подобряват качеството на живота ни. Освен това ИИ е от ключово значение за преодоляването на нерешените световни предизвикателства в областта на здравеопазването и благосъстоянието, изменението на климата, устойчивите правни и демографски системи и т.н., изразени в целите на ООН за устойчиво развитие.

ИИ може да бъде от огромна полза за хората и обществото, но той води и до някои рискове, които трябва да получат подобаващо внимание. Предвид факта, че като цяло ползите от ИИ са по-големи от рисковете, трябва да гарантираме, че ще поемем по пътя, който **води до максимално увеличаване на ползите от ИИ при свеждане на рисковете от него до минимум.** За да гарантираме, че се движим в правилната посока, **е необходимо да предприемем ориентиран към човека подход към ИИ,** който неизменно да ни напомня, че разработването и използването на ИИ не трябва да се разглеждат като самоцел, а като средство за повишаване на благосъстоянието на човека. **Наш ориентир ще бъде надеждният ИИ**, тъй като хората ще могат да се възползват уверено и напълно от предимствата, които осигурява ИИ, само ако могат да се доверят на технологиите.

Надеждният ИИ има **две основни характеристики**: 1) той трябва да зачита основните права, приложимата нормативна уредба и основните принципи и ценности, което гарантира неговата **„етично обоснована цел“**, и 2) той трябва да е **технически стабилен** и надежден, тъй като и при най-добри намерения, ако технологията не е овладяна, е възможно неумишлено да бъдат причинени вреди.

Ето защо в настоящите насоки се установява **рамка за надежден ИИ**:

- **в глава I** се разглежда **осигуряването на етично обоснованата цел на ИИ**, като се определят основните права, принципи и ценности, с които той трябва да е съобразен;
- въз основа на тези принципи в **глава II** са представени **насоки относно създаването** на надежден ИИ, като се разглеждат едновременно етично обоснованата цел и техническата стабилност. За целта са изброени изискванията за надежден ИИ и е включен преглед на техническите и нетехническите методи, които могат да бъдат използвани за създаването му;
- и накрая, **в глава III** е представена **практическата приложимост** на изискванията, като е даден конкретен, но неизчерпателен списък за оценяване на надеждния ИИ, който може да бъде адаптиран към отделните случаи.

За разлика от други документи по етичните аспекти на ИИ, целта на настоящите насоки не е съставянето на поредния списък на основните стандарти и принципи на ИИ, а да се предоставят

съвети за конкретното им прилагане и използване в системи с ИИ. Тези съвети са представени на три етапа с различна степен на теоретичност — глава I е най-теоретична (основни права, принципи и ценности), а глава III — най-конкретна (списък за оценяване).

Насоките са предназначени за всички **заинтересовани страни, които разработват, внедряват или използват ИИ**, в т.ч. дружества, организации, научни изследователи, публични служби, институции, физически лица или други субекти. В окончателната им редакция ще бъде представен механизъм, позволяващ на заинтересованите страни да ги подкрепят доброволно.

Важно е да се отбележи, че целта на насоките не е да се замени която и да е от провежданите политики или от прилаганите правни инструменти (по които ЕГВР ИИ ще работи по линия на втория документален резултат от своята дейност — препоръките относно политиката и инвестициите, които трябва да бъдат представени през май 2019 г.), нито да се възпрепятства въвеждането им. Нещо повече — насоките трябва да се разглеждат като динамичен документ, който трябва да се актуализира редовно, за да се гарантира тяхната актуалност спрямо развитието на технологиите и нашите познания за тях. Идеята е този документ да послужи като отправна точка за дискусия на тема **„надежден ИИ с марка „произведено в Европа“**.

Европа ще може да популяризира своя съобразен с етиката подход към ИИ едва когато стане конкурентоспособна в тази област в световен мащаб, но **само отчитането на етичните аспекти в областта на ИИ ще позволи постигането на отговорна конкурентоспособност**, тъй като по този начин ще се спечели доверието на потребителите и ще се улесни по-широкото навлизане на ИИ. Целта на настоящите насоки не е да се възпрепятстват иновациите в областта на ИИ в Европа, а етиката да се използва като източник на вдъхновение за разработването на единствена по рода си марка ИИ, чиято цел е да бъде от полза и да осигурява закрила както на отделните лица, така и на обществото. По този начин Европа ще може да заеме водещи позиции в сферата на авангардния, сигурен и етичен ИИ. Европейските граждани ще могат да се възползват изцяло от ползите, които носи ИИ, единствено ако не бъде накърнено доверието им в него.

И накрая, ако не се ограничаваме до европейското измерение, с настоящите насоки имат за цел **да подхранят анализа и дебата** относно етична рамка за ИИ в **световен план**.

НАСОКИ

Във всяка глава се предоставят насоки за създаването на надежден ИИ, предназначени за всички заинтересовани страни, които разработват, внедряват или използват ИИ. Тези насоки са обобщени по-долу:

Глава I: Ключови насоки за гарантиране на етично обоснованата цел:

- Уверете се, че ИИ е **ориентиран към човека**: ИИ трябва да се разработва, внедрява и използва с идеята за **етично обоснована цел**, т.е. той трябва да се основава на и да отразява основните права, обществените ценности и етичните принципи, които се свеждат до *действието в най-добрия интерес на ползвателите* (прави добро), *непричиняването на вреда* (не вреди), *автономността на хората*, *справедливостта* и *обяснимостта*. Това е от решаващо значение за работата по създаването на **надежден ИИ**;
- Изхождайте от основните права, етичните принципи и ценностите, за да направите оценка на предполагаемите въздействия на ИИ върху човешките същества и върху общото благо. Обърнете **особено внимание** на случаите, в които са засегнати повече **уязвими групи**, като например деца, лица с увреждания и малцинства, или на случаите на **несъразмерност на**

правомощията или на достъпа до информация, като например между работодатели и служители или предприятия и потребители;

- Задължително имайте предвид, че макар и да носи значителни ползи за отделните хора и за обществото, ИИ може да има и отрицателно въздействие. Обръщайте специално внимание на особено проблемните области.

Глава II: Ключови насоки за създаване на надежден ИИ:

- Включете **изискванията за надежден** ИИ **още от най-ранния етап на проектиране**: отчетност, управление на данните, универсален дизайн, управление на автономността на ИИ (контрол от страна на човека), недискриминация, зачитане на автономността на хората, зачитане на неприкосновеността на личния живот, стабилност, безопасност и прозрачност.
- Обмислете възможността за използване на технически и нетехнически методи, с които да гарантирате изпълнението на посочените изисквания в системата с ИИ. Освен това имайте предвид тези изисквания и при сформирането на екипа, който ще работи по системата, изграждането на самата система, средата за изпитване и потенциалните приложения на системата.
- Предоставете — недвусмислено и по своя инициатива — информация на заинтересованите страни (клиенти, служители и т.н.) за възможностите и ограниченията на системата с ИИ, което ще им позволи да си изградят реалистични очаквания. Осигуряването на **проследимост** на системата с ИИ е от решаващо значение в това отношение.
- Направете надеждния ИИ **част от организационната култура** и предоставете информация на заинтересованите страни за неговата роля при проектирането и използването на системи с ИИ. Надеждният ИИ може да бъде включен и в хартите за професионална етика или кодексите за поведение.
- Осигурете участието и **приобщаването на заинтересованите страни** на етапа на проектирането и разработването на системата с ИИ. Освен това осигурете **разнообразие** при сформирането на екипите, които разработват, внедряват и изпитват продукта.
- Постарайте се да **осигурите одитируемост** на системите с ИИ, особено в условия или случаи от особена важност. Доколкото е възможно, проектирайте системата си по такъв начин, че отделните решения да могат да се проследяват до различните входящи елементи: данни, предварително обучени модели и т.н. Освен това установете **методи за обяснение** на системата с ИИ.
- Осигурете наличието на специален процес за **механизмите за осигуряване на отчетност**.
- Предвидете **обучения** и направете необходимото ръководните кадри, разработчиците, потребителите и работодателите да бъдат осведомени и да преминат обучение в областта на надеждния ИИ.
- Имайте предвид, че може да са налице принципни противоречия между различните цели (прозрачността може да доведе до към злоупотреби, а откриването на склонности и предприемането на действия за тяхното коригиране може да бъде в разрез със закрилата на неприкосновеността на личния живот). Съобщавайте и документирайте компромисните решения в тези области.
- Насърчавайте научните изследвания и иновациите, за да съдействате за по-нататъшното изпълнение на изискванията за надежден ИИ.

Глава III: Ключови насоки за оценяване на надеждния ИИ:

- Приемете **списък за оценяване** на надеждния ИИ при разработването, внедряването или използването на ИИ и го адаптирайте към конкретния случай, в който се използва системата.

- Имайте предвид, че списъкът за оценяване никога няма да е изчерпателен, както и че осигуряването на надеждност не означава поставяне на отметки в кутийки, а представлява непрекъснат процес на установяване на изисквания, оценяване на решения и гарантиране на по-добри резултати през целия жизнен цикъл на системата с ИИ.

Настоящите насоки представляват част от визия, при която се възприема ориентиран към човека подход към изкуствения интелект, което ще осигури възможност на Европа да се превърне във водещ световен новатор в областта на етичния, сигурен и авангарден ИИ. Целта е да се улесни и да се позволи създаването на **„надежден ИИ с марка „произведено в Европа“**, който ще бъде в услуга на европейските граждани.