|  |  |
| --- | --- |
|  | **Софийски университет „Св. Кл. Охридски”**  Факултет по математика и информатика  *Магистърска програма „Изкуствен интелект”* |

**Предмет: Отговорен изкуствен интелект: етични, правни, социални и икономически аспекти**

***Зимен семестър, 2019/2020 год.***

**Тема №4: „Демонстриране на надеждност при системи с изкуствен интелект“**

**Задача №2: Подготовка на задълбочена презентация**

*Автор:*

*Иван Иванов Иванов, фак. номер 26100*

декември, 2019

София

Съдържание

[Изисквания за надежден ИИ 3](#__RefHeading___Toc42_1062697866)

[1 Човешки фактор и надзор 4](#__RefHeading___Toc44_1062697866)

[2 Техническа стабилност и безопасност 5](#__RefHeading___Toc46_1062697866)

[3 Управление на данните и неприкосновеност на личния живот 6](#__RefHeading___Toc48_1062697866)

[4 Прозрачност 7](#__RefHeading___Toc50_1062697866)

[5 Многообразие, недискриминация и справедливост 8](#__RefHeading___Toc52_1062697866)

[6 Обществено и екологично благополучие 9](#__RefHeading___Toc54_1062697866)

[7 Отчетност 10](#__RefHeading___Toc56_1062697866)

# Изисквания за надежден ИИ

Съществува списък от седем изисквания описващи прилагането и създаването на надежден ИИ. Тези изисквания са приложими по отношение на различни заинтересовани страни, които участват в жизнения цикъл на системи с ИИ: разработчици, внедрители и крайни ползватели, както и широката общественост. Под разработчици разбираме тези, които изследват, проектират и/или разработват системи с ИИ. Под внедрители разбираме публични или частни организации, които използват системи с ИИ в рамките на своите бизнес процеси и предлагат продукти и услуги на други лица. Крайните ползватели са тези, които са ангажирани пряко или косвено със системи с ИИ. И накрая, широката общественост обхваща всички други, които са пряко или косвено засегнати от системите с ИИ.

При гарантирането на спазването на въпросните изисквания различните групи заинтересовани страни изпълняват различни роли:

* **разработчиците** следва да изпълняват и прилагат изискванията във връзка с процесите на проектиране и разработване.
* **внедрителите** следва да гарантират, че използваните от тях системи и продуктите и услугите, които предлагат, отговарят на изискванията.
* **крайните ползватели** и **широката общественост** следва да бъдат информирани за тези изисквания и да са в състояние да изискват тяхното спазване.

**Списък с изискванията за създаване на надежден ИИ:**

1. Човешки фактор и надзор.
2. Техническа стабилност и безопасност Включително устойчивост на атака и сигурност, резервен план и обща безопасност, точност, надеждност и възпроизводимост.
3. Управление на данните и неприкосновеност на личния живот. Включително зачитане на правото на неприкосновеност на личния живот, качество и цялост на данните и достъп до данни.
4. Прозрачност включително проследимост, обяснимост и комуникация.
5. Многообразие, недискриминация и справедливост включително избягване на несправедливи предубеждения, достъпност и универсален дизайн и участие на заинтересованите страни
6. Обществено и екологично благополучие включително устойчивост и дейности, съобразени с околната среда, социално въздействие, общество и демокрация
7. Отчетност включително и свеждане до минимум и докладване на отрицателно въздействие, компромиси и правна защита.

# Човешки фактор и надзор

Системите с ИИ трябва да подпомагат автономността на хората и вземането на решения, както е определено от принципа на зачитане на автономността на хората. Това изисква системите с ИИ да спомагат за изграждане на демократично, процъфтяващо и справедливо общество чрез подкрепа на дейността на ползвателя и насърчаване на основните права, както и да позволяват човешки надзор.

**Основни права.** Както много технологии, системите с ИИ могат едновременно да зачитат и да възпрепятстват основните права. Те могат да бъдат от полза за хората, като например им помагат да проследяват личните си данни или като увеличават достъпността до образование и по този начин подкрепят правото им на образование. Предвид обхвата и капацитета на системите с ИИ обаче, те могат да оказват и отрицателно въздействие върху основните права. В ситуации, в които съществуват рискове като тези, следва да се извърши оценка на въздействието върху основните права. Тя следва да се извърши преди тяхното разработване и да включва оценка дали тези рискове могат да бъдат намалени или обосновани като необходими в демократичното общество, за да се зачитат правата и свободите на другите. Освен това следва да бъдат въведени механизми за получаване на външна обратна информация относно системи с ИИ, които потенциално нарушават основните права.

**Човешки фактор.** Ползвателите следва да са в състояние да вземат информирани автономни решения по отношение на системи с ИИ. Те следва да получат знания и инструменти, за да разбират и да си взаимодействат в задоволителна степен със системите с ИИ и когато е възможно, да имат възможност да направят задоволителна самооценка или да се противопоставят на системата. Системите с ИИ следва да подкрепят лицата да правят по-добър и по-добре информиран избор в съответствие с техните цели. Понякога системите с ИИ могат да бъдат внедрени, за да оформят и да окажат въздействие върху човешкото поведение чрез механизми, които могат да бъдат трудни за откриване, тъй като могат да включват подсъзнателни процеси, включително различни форми на нечестно манипулиране, заблуда, подбиране и уговаряне, като всички те могат да застрашат личната автономност. Общият принцип за автономността на ползвателя трябва да заема централно място във функционалността на системата. Ключ към това е правото на ползвателя да не бъде обект на решение, което се основава единствено на автоматизирана обработка, когато това води до правни последици за ползвателите или ги засягат съществено по подобен начин.

**Човешки надзор.** Човешкият надзор спомага да се гарантира, че дадена система с ИИ не накърнява автономността на хората и не причинява други неблагоприятни последици. Надзорът може да бъде постигнат чрез механизми за управление, като например прилагане на подход, при който човекът участва в процеса (HITL) или го наблюдава (HOTL) или контролира (HIC). HITL се отнася до възможността за човешка намеса във всеки цикъл за вземане на решение на системата, което в много случаи не е нито възможно, нито желателно. HOTL се отнася до възможността за човешка намеса по време на цикъла на проектиране на системата и наблюдение на функционирането на системата. HIC се отнася до възможността да се наблюдава цялостната дейност на системата с ИИ (включително нейното по-широко икономическо, обществено, правно и етично въздействие) и възможността да се взема решение кога и по какъв начин да се използва системата във всяка конкретна ситуация. Това може да включва решението да не се използва система с ИИ в конкретна ситуация, да се установят равнищата на лична преценка от страна на човека в хода на използване на системата или да се гарантира възможността за отмяна на решение, взето от системата. Освен това трябва да се гарантира, че публичните правоприлагащи органи са в състояние да упражняват надзор съобразно своя мандат. Механизмите за осъществяване на надзор могат да са необходими в различна степен за подкрепа на други мерки за безопасност и контрол в зависимост от областта на приложение на системата с ИИ и потенциалния риск. При равни други условия, колкото е по-малък надзорът върху дадена система с ИИ, който може да се упражнява от човек, се изисква толкова по-задълбочено изпитване и по-строго управление.

# Техническа стабилност и безопасност

Основен компонент за постигането на надежден ИИ е техническата стабилност, която е тясно свързана с принципа на предотвратяване на вреди. Техническата стабилност изисква системите с ИИ да бъдат разработени с превантивен подход по отношение на рисковете и по такъв начин, че да имат надеждно поведение съгласно предвиденото, като същевременно свеждат до минимум непреднамерените и неочакваните вреди и предотвратяват неприемливи вреди. Това следва да се прилага и по отношение на потенциални промени в работната им среда или присъствието на други действащи лица (хора и изкуствен интелект), които могат да си взаимодействат със системата по неблагоприятен начин. Освен това следва да се гарантира физическата и психическата неприкосновеност на хората.

**Устойчивост на атака и сигурност.** Като всички софтуерни системи и системите с ИИ следва да бъдат защитени срещу уязвимост, която може да позволи да бъдат използвани от противници, напр. хакерска атака. Атаките могат да са насочени към данните (заразяване на данни), модела (изтичане на модел) или към основната инфраструктура — както софтуер, така и хардуер. Ако е атакувана система с ИИ, напр. при вражески атаки, данните, както и поведението на системата, могат да бъдат променени, което ще накара системата да вземе различни решения или ще я накара да се изключи напълно. Системите и данните могат да бъдат увредени и от злонамерена намеса или от настъпването на непредвидени ситуации. Недостатъчните процеси, свързани със сигурността, също могат да доведат до погрешни решения или дори до физически вреди. За да се счита, че системите с ИИ са сигурни, под внимание следва да бъдат взети възможните нежелани приложения на ИИ (например приложения с двойна употреба) и потенциална злоупотреба със система с ИИ от злонамерени лица, като следва да се предприемат мерки за тяхното предотвратяване и смекчаване.

**Резервен план и обща безопасност.** В системите с ИИ следва да са включени предпазни мерки, които дават възможност за изпълнение на резервен план при възникване на проблеми. Това може да означава, че системи с ИИ могат да преминават от статистическа към базирана на правила процедура или че могат да поискат участието на човек оператор преди да продължат действието си39. Трябва да се гарантира, че системата ще изпълни това, което се очаква от нея, без да вреди на живите същества или на околната среда. Това включва свеждане до минимум на нежеланите последици и грешки. Освен това, в различните области на приложение следва да бъдат въведени процеси за разясняване и оценка на потенциалните рискове, свързани с използването на системи с ИИ. Нивото на необходимите мерки за безопасност зависи от степента на породения от система с ИИ риск, който на свой ред зависи от възможностите на системата. Когато може да се предвиди, че процесът на разработване или самата система ще породи конкретни значителни рискове, е от съществено значение мерките за сигурност да бъдат активно разработени и изпитани.

**Точност.** Точността се отнася до способността на система с ИИ да извършва правилни преценки, например да класифицира правилно информация в подходящите категории или способността ѝ да изготвя точни прогнози, препоръки или да взема правилни решения въз основа на данни или модели. Един подробен и добре организиран процес на разработване и оценка може да подкрепи, смекчи и да коригира непреднамерените рискове от неточни прогнози. В случай че не могат да се избегнат случайни неточни прогнози, е важно системата да посочи вероятността от появата на тези грешки. Високата степен на точност е от особено решаващо значение в ситуации, в които системата с ИИ засяга пряко човешкия живот.

**Надеждност и възпроизводимост.** От решаващо значение е резултатите от системите с ИИ да са възпроизводими, както и надеждни. Надеждна система с ИИ е тази, която работи правилно с набор от различни входящи данни и в различни ситуации. Това е необходимо за упражняване на контрол върху система с ИИ и за предотвратяване на непреднамерени вреди. Възпроизводимостта описва дали при даден експеримент с ИИ се наблюдава едно и също поведение, когато той се повтаря при едни и същи условия. Това дава възможност на учените и на създателите на политики да опишат точно какво правят системите с ИИ. Файловете за възпроизвеждане могат да улеснят процеса на изпитване и възпроизвеждане на поведения.

# Управление на данните и неприкосновеност на личния живот

Тясно свързана с принципа на предотвратяване на вреди е неприкосновеността на личния живот — основно право, което е особено засегнато от системите с ИИ. Предотвратяването на вредите по отношение на неприкосновеността на личния живот налага също подходящо управление на данните, което обхваща качеството и целостта на използваните данни, тяхното значение спрямо областта, в която системите с ИИ ще бъдат внедрени, протоколите им за достъп и способността за обработка на данните по начин, който защитава неприкосновеността на личния живот.

**Неприкосновеност на личния живот и защита на данните.** Системите с ИИ трябва да гарантират неприкосновеността на личния живот и защитата на данните през целия жизнен цикъл на системата41 . Това включва първоначално подадената от ползвателя информация, както и събраната за ползвателя информация в хода на взаимодействието му със системата (напр. изходящи данни, генерирани от системата с ИИ за конкретни ползватели или начините, по които ползвателите отговарят на конкретни препоръки). Цифровите записи на човешкото поведение могат да позволят на системите с ИИ да правят заключения не само за предпочитанията на хората, но и за тяхната сексуална ориентация, възраст, пол, религиозни или политически възгледи. За да могат лицата да се доверят на процеса на събиране на данни, трябва да се гарантира, че събраните във връзка с тях данни няма да бъдат използвани незаконосъобразно или несправедливо като средство за дискриминация срещу тях.

**Качество и цялост на данните.** Качеството на използваните набори от данни е от първостепенно значение за работата на системите с ИИ. Когато се събират данни, те могат да отразяват социално изградени предубеждения, неточности и грешки. Тези проблеми трябва да бъдат разрешени преди обучението с всеки даден набор от данни. Освен това трябва да се гарантира и целостта на данните. Въвеждането на злонамерени данни в система с ИИ може да промени нейното поведение, особено при самообучаващи се системи. Процесите и наборите от данни трябва да се изпитват и документират на всеки етап - планиране, обучение, изпитване и внедряване. Това следва да се прилага и за системи с ИИ, които не са разработени в рамките на дадена организация, а са придобити от друго място.

**Достъп до данни.** Във всяка организация, която работи с данни на физически лица (независимо от това дали става дума за ползвател на системата или не), следва да бъдат въведени протоколи за данни, уреждащи достъпа до данните. В тези протоколи следва да се посочва кой има достъп до данните и при какви обстоятелства. Достъпът до тях следва да има само надлежно квалифициран персонал, който е компетентен и се нуждае от достъп до данните на физическите лица.

# Прозрачност

Това изискване е тясно свързано с принципа на обяснимост и обхваща прозрачността на елементи, свързани със система с ИИ: данните, системата и бизнес моделите.

**Проследимост.** Наборите от данни и процесите, които са довели до решението на системата с ИИ, включително тези за събирането и етикетирането на данни, както и за използваните алгоритми, следва да бъдат документирани в съответствие с възможно най-добрия стандарт, за да се даде възможност за проследимост и повишаване на прозрачността. Това се прилага и по отношение на решения, взети от системата с ИИ. Това позволява установяване на причините, поради които решението, взето с ИИ, е погрешно, което от своя страна може да спомогне за предотвратяване на бъдещи грешки. Следователно прозрачността улеснява одитируемостта и обяснимостта.

**Обяснимост.** Обяснимостта се отнася до способността да се разясняват както техническите процеси на система с ИИ, така и свързаните с тях решения, взети от хора (напр. области на прилагане на система с ИИ). Техническата обяснимост изисква взетите от система с ИИ решения да могат да бъдат разбрани и проследени от хората. Освен това може да се наложи да се постигне компромис между повишаване на обяснимостта на дадена система (което може да намали нейната точност) или повишаване на нейната точност (в замяна на обяснимостта). Всеки път, когато една система с ИИ оказва значително въздействие върху живота на хората, следва да е възможно да се изисква подходящо обяснение на процеса на вземане на решения на системата с ИИ. Такова обяснение следва да бъде навременно и адаптирано към нивото на експертни знания на въпросната заинтересована страна (напр. неспециалист, регулаторен орган или изследовател). Освен това следва да бъдат предоставени обяснения за степента, до която дадена система с ИИ въздейства върху и оформя организационния процес на вземане на решения, за проектантските решения при създаване на системата и за мотивите за въвеждането ѝ (по този начин ще се гарантира прозрачността на съответния бизнес модел).

**Комуникация.** Системите с ИИ не следва да се представят пред потребителите като хора. Хората имат правото да бъдат информирани, че си взаимодействат със система с ИИ. Това означава, че системите с ИИ трябва да бъдат лесноразпознаваеми. Също така, когато е необходимо, следва да се предвиди възможност за вземане на решение срещу това взаимодействие в полза на човешкото взаимодействие, за да се гарантира спазването на основните права. Освен това възможностите и ограниченията на системата с ИИ следва да се съобщават по подходящ начин на професионалистите в областта на ИИ или на крайните ползватели, участващи в използването ѝ. Това би могло да обхване съобщаването на степента на точност на системата с ИИ, както и на нейните ограничения.

# Многообразие, недискриминация и справедливост

За да постигнем надежден ИИ, трябва да предоставим възможност за включване и многообразие през целия жизнен цикъл на система с ИИ. Освен че се взема внимание и участието на всички засегнати заинтересовани страни по време на процеса, това води също до гарантиране на равен достъп чрез процеси на приобщаващо проектиране, както и равно третиране. Това изискване е тясно свързано с принципа на справедливост.

**Избягване на несправедливи предубеждения.** Наборите от данни, използвани от системите с ИИ (както за обучение, така и при експлоатация), могат да бъдат засегнати от включването на непреднамерени исторически предубеждения, непълнота и лоши модели на управление. Продължаването на подобни предубеждения може да доведе до непреднамерени преки и непреки предразсъдъци и дискриминация42 срещу определени групи хора, потенциално изостряне на предразсъдъците и маргинализация. Преднамереното използване на предубеждения (на потребители) или участието в нелоялна конкуренция, като например хомогенност на цените чрез тайни споразумения или непрозрачен пазар, също могат да причинят вреда43. Когато е възможно, на етапа на събиране следва да се премахнат лесноразпознаваеми и дискриминационни предубеждения. Начинът, по който се разработват системите с ИИ (напр. програмиране на алгоритми), също може да бъде засегнат от несправедливи предубеждения. На това може да се противодейства чрез въвеждането на надзорни процеси за анализиране и разглеждане по ясен и прозрачен начин на целта, ограниченията, изискванията и решенията на системата. Освен това следва да се насърчава и наемането на лица от различни среди, култури и области, което може да осигури многообразие от мнения.

**Достъпност и универсален дизайн.** Системите следва да са ориентирани към потребителя и да са проектирани по начин, който позволява на всички хора да използват продукти или услуги с ИИ, независимо от своята възраст, пол, способности или характеристики, особено в областите, включващи отношения между предприятия и потребители. От особено значение е достъпността до тази технология за лица с увреждания, които са представени във всички обществени групи. Системите с ИИ не следва да използват универсален подход по отношение на всички и следва да зачитат принципите за универсален дизайн44, насочени към възможно най-широк кръг ползватели, като се следват съответните стандарти за достъпност45 . Това ще даде възможност за справедлив достъп и активно участие на всички хора в съществуващи и възникващи човешки дейности с участието на компютри, както и по отношение на помощни технологии.

**Участие на заинтересованите страни.** За да се разработят надеждни системи с ИИ, е препоръчително да се провеждат консултации със заинтересовани страни, които могат пряко или косвено да бъдат засегнати от системата в рамките на жизнения ѝ цикъл. Полезно е да се иска редовна обратна информация дори и след внедряването и да се създават по-дългосрочни механизми за участие на заинтересованите страни, например чрез осигуряване на информация, консултации и участие на работниците през целия процес на въвеждане на системите с ИИ в организациите.

# Обществено и екологично благополучие

В съответствие с принципите на справедливост и предотвратяване на вредите широката общественост, другите същества с усещания и околната среда също следва да се разглеждат като заинтересовани страни през целия жизнен цикъл на ИИ. Устойчивостта и екологичната отговорност на системите с ИИ следва да бъдат насърчавани, както и изследванията, насочени към решения с използване на ИИ в области от глобално значение, като например целите за устойчиво развитие. В идеалния случай ИИ следва да се използва в полза на всички хора, включително бъдещите поколения.

**Устойчив и екологичен ИИ.** Системите с ИИ имат потенциала да помогнат за справяне с някои от най- неотложните обществени проблеми, макар че трябва да се гарантира, че това ще се случи по възможно най-екологичния начин. Процесът на разработване, внедряване и използване на системата, както и цялата верига за доставка, следва да се оценяват в тази връзка, напр. чрез критична проверка на използването на ресурсите и потреблението на енергия по време на обучението, като се правят избори, причиняващи по-малка вреда. Мерките за гарантиране на съобразяване с околната среда на цялата верига за доставки на системата с ИИ следва да се насърчават.

**Социално въздействие.** Повсеместното излагане на социални системи с ИИ във всички области на нашия живот (било то в образованието, работата, грижите или забавленията) може да промени нашата представа за социалния фактор или да повлияе на социалните ни взаимоотношения и привързаност. Макар че системите с ИИ могат да се използват за подобряване на социалните умения , те могат да допринесат и за тяхното влошаване. Това може да засегне и физическото и психическото благосъстояние на хората. Следователно последствията от тези системи трябва да бъдат внимателно наблюдавани и разглеждани.

**Общество и демокрация.** Освен оценката на въздействието на разработването, внедряването и използването на система с ИИ върху хората, това въздействие следва да бъде оценено и от обществена гледна точка, като се вземе предвид неговото въздействие върху институциите, демокрацията и обществото като цяло. Използването на системи с ИИ следва да се разглежда внимателно, особено в ситуации, свързани с демократичния процес, включително не само при вземането на политически решения, но и при провеждането на избори.

# Отчетност

Изискването за отчетност допълва горепосочените изисквания, като е тясно свързано с принципа на справедливост. То налага въвеждането на механизми за гарантиране на отговорността и отчетността на системите с ИИ и резултатите от тях както преди, така и след тяхното въвеждане.

**Одитируемост.** Одитируемостта включва въвеждането на оценката на алгоритми, данни и процеси на проектиране. Това не означава задължително, че информацията за бизнес модели и интелектуална собственост, свързана със системата с ИИ, трябва винаги да бъде свободно достъпна. Оценката от вътрешни и външни одитори и наличието на подобни доклади за оценка могат да допринесат за надеждността на технологиите. При приложения, засягащи основните права, включително приложения от критично значение за безопасността, системите с ИИ следва да са в състояние да бъдат одитирани независимо.

**Свеждане до минимум и докладване на отрицателни въздействия.** Както възможността да се докладва за действия или решения, които допринасят за определен системен резултат, така и възможността да се отговаря на последствията от този резултат трябва да бъдат гарантирани. Установяването, оценката, докладването и свеждането до минимум на потенциални отрицателни въздействия на системи с ИИ е от съществено значение за тези, които са пряко(косвено) засегнати. Необходимо е да се осигури защита на лицата, които подават сигнали, неправителствените организации, синдикалните организации или други организации, когато докладват за основателни опасения във връзка с базирана на ИИ система. Използването на оценки на въздействието (напр. тестова кибератака с цел подобряване на сигурността или форми на алгоритмична оценка на въздействието) както преди, така и по време на разработването, внедряването и използването на системи с ИИ може да бъде полезно за свеждането до минимум на отрицателното въздействие. Тези оценки трябва да бъдат пропорционални на риска, който пораждат системите с ИИ.

**Компромиси.** При изпълнение на горепосочените изисквания между тях могат да възникнат противоречия, което може да доведе до неизбежни компромиси. Такива компромиси следва да се разглеждат по рационален и методичен начин съгласно степента на развитие на техниката. От това следва, че трябва да се идентифицират съответните интереси и ценности, засегнати от системата с ИИ и, ако възникне конфликт, компромисите между тях трябва да бъдат изрично признати и оценени от гледна точка на риска, който пораждат за етичните принципи, включително основните права. В ситуации, в които не могат да се установят приемливи от етична гледна точка компромиси, разработването, внедряването и използване на система с ИИ не следва да се извършва по този начин. Всяко решение за това какъв компромис трябва да се направи трябва да бъде обосновано и надлежно документирано. Лицето, вземащо решенията, трябва да носи отговорност за начина, по който се прави подходящия компромис, както и непрекъснато да преразглежда целесъобразността на взетото решение, за да гарантира, че в системата могат да бъдат направени необходимите промени, когато това е необходимо.

**Правна защита.** При възникване на несправедливи неблагоприятни последици следва да бъдат предвидени достъпни механизми за гарантиране на подходяща правна защита. Knowing that redress is possible when things go wrong is key to ensure trust. Специално внимание следва да се обърне на уязвимите лица или групи.