Етика
Наука
Образование

Издание
на „Лаборатория по приложна академична етика“ и
катедра „Философски науки“

ISSN 3033-005X (Print)
3033-0068 (Online)

Година II, Брой 2, 2024

DOI: 10.54664/EKZC7283

АКАДЕМИЧНА ПОЧТЕНОСТ: МЕЖДУ НОРМАТИВНА УРЕДБА И ВЪЗПРИЯТИЯ ЗА ЕТИКАТА НА ИЗПОЛЗВАНЕТО НА ГЕНЕРАТИВЕН ИИ

Мария Чанкова, Ирена Василева

ACADEMIC INTEGRITY: BETWEEN REGULATION AND PERCEPTIONS OF THE ETHICS OF USING GENERATIVE AI

Mariya Chankova, Irena Vassileva

Abstract: The article deals with the issues of using generative AI tools and Large Language Models in education and research from two viewpoints. First, it presents an overview of existing regulatory documents concerning AI tools at the EU level and the lack thereof in Bulgaria. Then it focuses on users' perception and understanding of AI tools through an overview of available video materials of instructional character, with a special focus on (non)existing ethical concerns, as well as software producers' perspectives on the affordances and weaknesses of the technology. We conclude that there is an urgent need to create guidelines/rules for the use of AI systems in relevant fields and recommend that the Ministry of Education and Science create guidelines to be implemented by educational and research institutions and publishers.

Keywords: Academic integrity; Generative AI in education and research; Regulations; Perceptions

Увод

С предоставянето за свободно ползване на ИИ чатбот през ноември 2022 г. в академичното пространство се подновяват дискусиите относно ползите и вредите от тази технология за различните области на научните изследвания. Възможностите на технологията да генерира успешна имитация на човешка реч и капацитетът ѝ да се самообучава, и следователно, потенциалът ѝ да се усъвършенства с времето (Bender et al. 2021) е в основата на дискусиите. Оценката на постигнатото от OpenAI до момента не е еднозначна: Чомски, например, за когото полезността на разработките в сферата на ИИ е в потенциала им да помогнат да се разбере как функционира човешкото съзнание, определя системи от типа на ChatGPT като „безполезни“ и „игра със засукани играчки“ („playing with fancy toys“)[1], освен за студента, който иска да препише на изпит. Marcus определя системите като „autocomplete on steroids“ (автоматично довършване на стероиди) и изтъква потенциалната заплаха за информационната чистота на екосистемите от съдържание, в които съществуват човешките общества. Според него, наводняването на интернет пространството с фалшива, генерирана дезинформация, е въпрос на време и представлява сериозна заплаха за

[1] Всички преводи в статията са на авторите.

демократичните процеси[2]. Основната линия на критика от страна на компютърните специалисти остава стабилността на ИИ системите, като например стабилност на получените резултати (Marcus & Davis 2020), поради което изглежда, че постиженията на ИИ не изпълняват обещанието си за технологична революция.

В настоящата статия ще разгледаме някои проблеми, засягащи академичната почтеност и в частност етиката на използване на ИИ-базирани продукти в областта на научната продукция и образованието, като се постараем да обхванем нормативната страна на въпроса (как и дали тези проблеми намират отражение в нормативната база на научните и образователни институции), както и комуникацията, протичаща в интернет пространството, която отразява какво търсят и предлагат потребителите (професионалисти, учени, студенти и др.) по тези въпроси.

Законодателни рамки

През март 2024 г. ЕП приема първия в света Законодателен акт за използване на изкуствения интелект[3], където „Ключова характеристика на системите с ИИ е тяхната способност да правят изводи. Тази способност да се правят изводи се отнася до процеса на получаване на резултати, например прогнози, съдържание, препоръки или решения.„ В документа се определят три категории системи за изкуствен интелект:

- С неприемлив риск, чието използване се забранява в рамките на ЕС;
- С висок риск, чието използване се регулира от закона;
- С минимален риск, чието използване не се регулира.

В първата група попадат основно приложения, които могат да доведат до нарушаване на човешките права, представляват заплаха за здравето и безопасността на хората. Без да навлизаме в подробности, в контекста на настоящата статия трябва да отбележим, че като една от областите с висок риск е определено образованието на всички негови нива. От една страна, в документа се насърчава внедряването на системи с ИИ с цел придобиване на необходимите дигитални умения и компетентности с оглед на бъдещото активно участие на учащите се в икономиката и обществените процеси. От друга страна се посочват някои рискове, като: „определяне на достъпа или приема, за разпределяне на лица в институции или програми на всички нива на образованието и професионалното обучение, за оценяване на учебните резултати, […] или за наблюдение и откриване на забранено поведение на ученици и студенти по време на тестове, следва да се класифицират като високорискови системи с ИИ, тъй като могат да определят образователния и професионалния път на дадено лице и така да повлияят на способността на това лице да осигурява прехраната си.“(Законодателен акт, стр. 54)

През 2022 г. ЕК публикува „Етични насоки за преподавателите относно използването на изкуствен интелект (ИИ) и на данни при преподаване и учене„[4], които намират приложение в България едва през февруари 2024 г., когато МОН лансира „Насоки за използване на изкуствен интелект в образователната система.“[5]

Законодателният акт на ЕП обаче се оказва недостатъчен за регулиране на използването на системи за ИИ в научните изследвания. Ето защо, в следствие на практически мигновеното отражение на предоставянето за свободно ползване на системата ChatGPT в края на 2022 г. върху учебната и изследователската работа на висшите учебни заведения и изследователските институции, както и на издателствата на научна литература, голяма част от тях се заемат с изготвянето на правила за използването на тази система, опитвайки се да предотвратят вече проявяващите се негативни последствия. Обобщавайки вече събрания опит, ЕК публикува през март 2024 г. „Living

[2] Chomsky и Marcus се изказват на форум, посветен на ИИ през февруари 2023 г. Web Summit.

[3] Версията на български език е достъпна на: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_BG.pdf

[4] https://op.europa.eu/bg/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1

[5] https://www.mon.bg/nfs/2024/02/nasoki-izpolzvane-ii_190224.pdf

guidelines on the responsible use of generative AI in research“[6] („Живи насоки за отговорното използване на генеративния ИИ в научните изследвания“). Насоките са наречени „живи“, тъй като подлежат на постоянно актуализиране в зависимост от получената обратна информация от всички заинтересовани страни и поради изключително бързото развитие на системите с ИИ. Тук само ще отбележим, че все още документът няма официален превод на български език.

В него се засягат три сфери, свързани с „отговорното използване на генеративния ИИ в научните изследвания“ (стр. 4), а именно: (1) препоръки за учените; (2) препоръки за изследователските организации и (3) препоръки за организациите, финансиращи научни изследвания. Създаването на документа от Форума на Европейското изследователско пространство е провокирано от непрекъснато роящите се насоки, указания, препоръки, правилници и др., изготвени от университети, научни и финансиращи организации, като целта е те да се организират в единен документ, който да улесни конкретизирането на мерките в зависимост от нуждите и спецификите в контекста на дадената страна. Същевременно се подчертава, че системите с ИИ могат да окажат негативно влияние върху интегритета на научните изследвания и отделните учени, като доведат до неетични научни практики и дезинформация. Докато някои от рисковете при използване на ИИ се дължат на несъвършенства в самите системи и предоверяването към тях, други се съдържат в тяхната некоректна съзнателно или несъзнателно употреба.

Ключовите принципи на Насоките се основават на предходни рамкови документи, включително и на Европейският етичен кодекс за почтеност на научните изследвания в неговото актуализирано издание от 2023 г.[7] Тези принципи по отношение на използването на ИИ са (стр. 7):

1. Надеждност – „Включва аспекти, свързани с проверката и възпроизвеждането на информацията, произведена от ИИ за научни изследвания“.

2. Честност – „Този принцип включва деклариране, че е използван генеративен ИИ“.

3. Уважение – „Отговорното използване на генеративен ИИ трябва да отчита ограниченията на технологията, нейното въздействие върху околната среда и нейните обществени ефекти […]. Това включва правилно управление на информацията, зачитане на неприкосновеността на личния живот, поверителността и правата на интелектуална собственост и правилно цитиране.

4. Отчетност – „Това включва отговорност за всички резултати, произведени от изследователя, подкрепени от идеята за човешка намеса и надзор“.

Препоръки в „Живите насоки за отговорното използване на генеративния ИИ в научните изследвания“

(1) Препоръки за учените

В тази част на Насоките се набляга основно на отговорността на учените за тяхната продукция: отговорност за верността съдържанието, произведено от ИИ, което изисква критичен подход към него, свързан с ограничените му възможности и породените от тях неточности, т.н. халюцинации и пристрастия. Твърдо се заявява, че: „Системите за ИИ не са нито автори, нито съавтори. Авторството предполага човешка намеса и отговорност, така че то е дело на изследователя“ (стр. 8).

На второ място се обръща внимание на прозрачното използване на ИИ, т.е. когато е налично такова, то да бъде ясно и точно описано и посочено в съответната публикация, като се взема предвид „Стохастичния (случаен) характер на генеративните инструменти за ИИ, при който има тенденцията да се произвежда различен изходен материал от един и същ входящ материал“ (стр. 8).

Третата точка засяга необходимостта да се спазват правилата за „неприкосновеност, поверителност и права на интелектуална собственост при споделяне на чувствителна или защитена информация с инструменти за изкуствен интелект“ (стр. 8), които включват не само политиките за защита на личната информация на ЕС, но и споделяне със системи с ИИ на текстове, образи,

[6] https://research-and-innovation.ec.europa.eu/document/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?pref Lang=fr

[7] https://allea.org/portfolio-item/european-code-of-conduct-2023/

данни и др. без съгласието на авторите, тъй като те биха могли да бъдат използвани за обучение на системите и това да доведе до неправомерното им по-нататъшно използване.

На четвърто място се указва задължението учените да съблюдават националното и международното законодателство във връзка със спазването на правата върху интелектуална собственост, тъй като съществува реален риск продукцията на ИИ да представлява плагиатство без указване на оригиналния източник.

Във връзка с всички указания се посочва, че учените трябва да са непрекъснато в крак с бързо развиващите се възможности на ИИ както по отношение на добрите практики за неговото използване, така и поради потенциалните рискове.

На последно, но не маловажно място се отбелязва, че трябва да се минимизира използването на ИИ при оценяване на работата на учените, например при писане на рецензии или оценки на проекти с цел финансирането им, или пък при кандидатстване за дадено работно място.

(2) Препоръки за изследователските организации

На първо място, организациите трябва да „Насърчават, ръководят и подкрепят отговорното използване на генеративен ИИ в изследователски дейности“ (стр. 10), като осигуряват обучения за неговото етично и правомерно използване и предоставят указания, свързани с гарантирането на спазването на европейските и международни законодателства.

Освен това организациите трябва да следят за използването и развитието на ИИ в самите тях с цел подсигуряване на своевременни насоки за работа с него.

Особено важна е третата препоръка, а именно: „Да се позовават на тези насоки за генеративен ИИ, или да ги интегрират в техните общи насоки за научни изследвания за добри изследователски практики и етика“ (стр. 10), тоест, те да бъдат включени в Етичните кодекси, или да бъдат изведени като отделни документи при условие за непрекъснато осъвременяване.

Последната препоръка е от технически характер и визира създаването на институционални облачни пространства, до които имат достъп единствено служителите, което ще повиши нивото на сигурност и поверителност на данните.

(3) Препоръки за организациите, финансиращи научни изследвания

Поради понякога съществените разлики в контекста, в който финансиращите организации функционират, препоръките в тази част са от най-общ характер. Те не се различават особено от тези за изследователските организации и за отделните учени, а по-скоро ги обединяват, и затова няма да се спираме на тях по-подробно. Трябва да се отбележи обаче, че водещи такива организации вече са създали собствени насоки, реагирайки на новите обстоятелства. Така например, Deutsche Forschungsgemeinschaft (DFG, Германска изследователска фондация), най-голямата организация за финансиране на научни изследвания в Германия, излиза със становище и указания за кандидатите още през септември 2023 г.[8]

Освен на гореспоменатия Европейски етичен кодекс за почтеност на научните изследвания, Насоките се основават и на Насоките относно етичните аспекти за надежден ИИ[9], публикувани през 2019 г. Въпреки, че се отнасят предимно до етичните принципи за създаване на системи с ИИ, се смята, че те могат да послужат като отправна точка и за техните ползватели. Принципите са систематизирани по следния начин (Насоки, стр. 12.):

„Четирите етични принципа за ИИ системите са:

1. уважение към човешката автономия;
2. предотвратяване на вреди;
3. справедливост;
4. обяснимост.

[8] https://www.dfg.de/resource/blob/289676/89c03e7a7a8a02409360299597 4832f9/230921-statement-executive-committee-ki-ai-data.pdf

[9] https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf

Тези етични принципи бяха използвани за разработването на следните седем оперативни ключови изисквания:

1. човешка намеса и надзор;
2. техническа издръжливост и безопасност;
3. поверителност и управление на данните;
4. прозрачност;
5. разнообразие, недискриминация и справедливост;
6. благосъстояние на околната среда и обществото;
7. отчетност“.

Както се вижда от гореизброените етични принципи, на практика всички те са взети под внимание при създаването на Насоките за отговорното използване на генеративния ИИ в научните изследвания.

Използване на ИИ в публикуването на научни изследвания

Въпреки че дотук дискутираните Насоки не адресират изрично процеса на публикуване на научни изследвания, би могло да се каже, че основните принципи, залегнали в тях, се отнасят и до тази област. Всъщност, след няколкото опита да се посочи ChatGPT като съавтор на статии, големите международни издателства веднага реагираха със създаването на указания за авторите. Ganjavi et al (2023) правят библиометричен анализ на съществуващите към този момент инструкции за авторите във връзка с използването на генеративен ИИ (GAI), GPT модели и ГЕМ сред стоте водещи издателства и списания и стигат до следните резултати: „Сред 100-те най-големи издатели 17% предоставят насоки за използването на GAI, от които 12 (70,6%) са сред първите 25 издатели. Сред първите 100 списания 70% са предоставили насоки относно GAI. От тези с насоки 94,1% от издателите и 95,7% от списанията забраняват включването на GAI като автор„ (стр. 2). Авторите обаче установяват, че се наблюдават съществени разлики между указанията по отношение на техния обхват, степента на разрешена употреба на GAI и начините за нейното указване, включително и между афилиирани издателства и списания. Ето защо те заключават, че: „Липсата на стандартизация натоварва авторите и заплашва да ограничи ефективността на тези разпоредби. Има нужда от стандартизирани насоки, за да се защити интегритета на научната продукция, тъй като популярността на GAI продължава да нараства„ (стр. 2).

Към настоящия момент повечето от големите издателства, а и не само те, заявяват, че се придържат към препоръките на COPE (Committee on Publication Ethics) за „Авторство и инструменти с ИИ“[10], където основните положения могат накратко да се обобщят като:

– Инструменти с ИИ не могат да се впишат като автори: понятието за авторство се конструира около отговорността за идеите, които се представят; авторите са отговорни за това, тяхната работа да не нарушава правата на трети страни;

– Използването на инструменти с ИИ трябва да е прозрачно и надлежно указано по подходящ начин, когато дадена разработка се предава за публикуване;

– Рецензентите не трябва да качват ръкописи в инструменти с ИИ, тъй като това може да наруши правилата за конфиденциалност и защита на личните данни.

Междувременно се появиха редица статии, засягащи различни аспекти на използването на ИИ в научни публикации. Те варират между опит да се използва ChatGPT като съавтор, където се повдигат съществени въпроси, свързани с понятията „оригиналност“, „авторство“, „плагиатство“ (Frye, 2023) и напълно отричане: „високотехнологично плагиатство“ и „начин да се избегне ученето“ (Chomsky on EduKitchen 2023). Тук няма да се спираме подробно на този въпрос, тъй като темата е огромна, в непрекъснато развитие и процес на доуточняване.

Регулация на използването на системи с ИИ в България

В Законодателния акт на ЕП се казва, че: „Всички заинтересовани страни, включително промишлените сектори, академичните среди, гражданското общество и организациите за стан-

[10] https://publicationethics.org/cope-position-statements/ai-author

дартизация, се насърчават да вземат предвид, когато е целесъобразно, етичните принципи за разработването на доброволни най-добри практики и стандарти“ (Законодателен акт, стр. 27).

В България засега съществуват изключително ограничен брой документи, отнасящи се до използването на ИИ, като в контекста на тази статия ще спомена два от тях. През октомври 2020 г. Министерство на транспорта, информационните технологии и съобщенията публикува „Концепция за развитието на изкуствения интелект в България до 2030 г.“[11], в която обаче се залагат единствено положителните аспекти на използването на ИИ във всички възможни области на обществения и стопански живот, без да се разглеждат евентуалните рискове. Този документ е публикуван преди старта на свободната употреба на ChatGPT и последвалите го подобни системи и не е актуализиран оттогава.

През февруари 2024 г. МОН публикува „Насоки за използване на изкуствен интелект в образователната система“[12], които са ограничени до средното образование и не обхващат висшето образование и науката. Проблематично в този текст е твърдението, че: „ГЕМ: Разполагат с разширено разбиране за контекста на дадено запитване, което им позволява да предоставят по-детайлни и нюансирани отговори„ (стр. 9). Както доказват редица изследвания, именно липсата на разбиране на контекста е едно от слабите места на ГЕМ (Големи езикови модели), което води до продукция на нерелевантни, изкривени или предубедени резултати. Тук под „контекст„ се разбира конкретната ситуация на комуникацията с всички нейни компоненти, а не просто предходния и последващия текст, понякога наричан в литературата „ко-текст“.

Освен това в забележка под линия на горното твърдение се допълва, че: „AI могат също да превеждат съдържание за отговора си на езика, на който са били запитани, без значение, че са обучавани със съдържание на друг език. Това подобрява значително възможността да се правят задълбочени изследвания на тема независимо от наличността на най-новата информация на български„ (пак там). Литературата по въпроса за използването на ГЕМ за целите на превода обаче, поне засега, ясно показва, че: „GPT моделите постигат много конкурентно качество на превода за езици с висок ресурс, докато имат ограничени възможности за езици с нисък ресурс“ (Hendy et al. 2023, 1; вж. също Peng et al. 2023, Gao et al. 2023). Българският език принадлежи към втория тип езици и едно сравнително изследване демонстрира слабостите на ChatGPT като инструмент за превод на фона на друг език с висок ресурс (немски) и популярните невронно базирани инструменти Google Translate и DeepL (Vassileva 2024, под печат).

Поне на този етап авторите на тази статия не успяха да намерят други общодостъпни официални документи, регулиращи използването на ИИ в образованието и науката. Същевременно, от юридическа гледна точка, Куманова и Даскалова (2024, 3) установяват, че:

„Системният анализ на нормативните актове, които регулират правото на образование на българските граждани и респективно, задълженията на образователните институции заключава, че няма нормативна уредба на използването на изкуствения интелект. Налице е празнота в правото като израз на пълно или частично отсъствие на правни норми (повели) по отношение на конкретни обществени отношения, които предполагат, допускат и изискват правна регламентация„.

Що се отнася до политиките на издателства и отделни списания, в „списъка на съвременните български научни издания, реферирани и индексирани в световноизвестни бази данни с научна информация“ на НАЦИД[13] фигурират 196 издания, като около половината от тях са престанали да бъдат реферирани към днешна дата. Точна статистика не е възможна, тъй като повечето от изданията нямат електронен достъп, но общият преглед показва, че много малка част имат специален раздел „Етика“, където обикновено се споменава, че издателството се придържа към насоките на Committee on Publication Ethics (COPE). Никъде не се споменава използването на инструменти с ИИ.

[11] https://www.mtc.government.bg/sites/default/files/koncepciyazarazvitienaiivbulgariyado2030.pdf

[12] https://www.mon.bg/nfs/2024/02/nasoki-izpolzvane-ii_190224.pdf

[13] https://randii.nacid.bg/register/search

Нагласи и възприятия на етичните проблеми от потребителите

Докато институциите пишат и разпространяват насоки за етично използване на генеративен ИИ в научната и образователната сфера, разработчиците на ИИ-базирани продукти се ръководят от комерсиален интерес и развиват дейност, която потенциално затруднява значително усилията на институциите да ограничат рисковете и влиянието на тези продукти. Трябва да се отбележи, че в етичен план съществува неоднозначност на нагласите и възприятията на тези рискове. Следва да се уточни също, че в текста продължаваме да използваме „ИИ“ за улеснение, макар да имаме предвид по-конкретно генеративен ИИ, базиран на Големи езикови модели.

Van Dis et al. (2023) правят следната прогноза след тестване на технологията:

Потенциални предимства	**Потенциални проблеми**
Ускоряване на процеса на иновациите	Влошаване на качеството и прозрачността на научните изследвания
Скъсяване на времето преди публикуване	Фундаментална промяна на автономността на учените
Помощ в писането на учени чужденци	Фактически грешки и „салата от думи“, произведена от чат бота (за повече справедливост и разнообразие на научните перспективи)
Разпространяване на дезинформация	

Изводът, който се налага след подобни констатации, е за важността на човешкия редактор, който задължително трябва да провери верността и истинността на генерираните твърдения. Van Dis et al. (2023) предлагат да се изработи система за отговорно и прозрачно използване на ИИ, което да помага за регулиране количествено и качествено на генерираната продукция, в частност тази, която се влива в научната литература, без да предлагат конкретни мерки, които биха могли да изграждат такава система.

Dwivedi et al. (2021) акцентират на качественото въздействие на ИИ системи върху профила на компетентностите, които ще бъдат важни за учения: без да подават пример за такива, авторите настояват, че една важна задача на учените от различни области е да се предвидят промените, за да се адаптират към променящите се нагласи и основни предположения, върху които се гради дадена научна област. Макар подобно предложение да изглежда на пръв поглед необосновано, в по-слабата му форма то заслужава внимание: в сферата на образованието (процес на обучение, изпитни формати, оценяване, и т.н.), както и в сферата на разпространяването на научна литература, потенциалът на ИИ да промени драстично някои дейности е очевиден (Perera & Lankathilaka 2023). Разбира се, този потенциал може да доведе и до отрицателни ефекти, включително закърняване и загуба на някои навици за учене и обработване на информация (Datoo & Siddiqui 2024). Много учени предвиждат неизбежната надпревара между генеративни ИИ системи и системи за разпознаване на текст, генериран чрез ИИ, като вместо само да „автоматизира някои повтарящи се и трудоемки дейности“, чат ботът върши цялата работа вместо човека (Van Dis et al. 2023, Marcus 2023). Този и подобни етични въпроси за правилното интегриране на системите на ИИ в сферите на образованието и на научната продукция занимават научната общност от известно време.

Foltýnek (2024) предлага следното обобщение на етичните проблеми, свързани с използването на ИИ, които потенциално могат да се появят в научноизследователския процес в различните му етапи:

– Етап на проучване на теоретико-методическата литература – недостоверни източници (predatory journals – псевдонаучни списания)

– Обобщение на теоретико-методическата литература С неточна или подвеждаща информация

– Обработване на данните – грешки при кодиране, при обработване
– Създаване на текст – халюцинации, предубеждения, пристрастие
– Разпространение на научните резултати – опростяване, стереотипиране

За всички етапи на научноизследователския процес са възможни проблеми, свързани със защита на личните данни, защита на интелектуалната собственост.

Foltýnek не смята, че са необходими специални регулации или добавки към етичните правила, за да се регулира етичното използване на ИИ, тъй като според него проблемите, които поставя ИИ, са аналогични на проблемите, които могат да възникнат пред изследователите по принцип. Той препоръчва внимателна проверка на генерирания текст и напомня, че именно ученият е отговорен за предлагането и разпространяването на (частично) генерирано съдържание, като подчертава необходимостта от обучение по въпросите за ограниченията на ИИ и етичното му използване.

Тук е редно да квалифицираме оценката му: въпреки че нищо фундаментално ново като заплаха не стои пред интегритета на научно-изследователския процес, темповете, с които пространството може да се изпълни с фалшива (и/или посредствена) научна продукция, са главозамайващи. Нещата се усложняват ако се имат предвид някои резултати, получени от Gao et al. (2023), които изследват разликите между анотации, писани от човек, и такива, генерирани от ИИ: става все по-трудно за човешки редактор да различи анотации, писани от човек от тези, генерирани от ИИ, като няма статистически значителна разлика между резултатите, получени от човек и от софтуер за детекция на генерирано съдържание. Анотациите, използвани в изследването, са от областта на медицината - една област, в която езикът и стилът имат второстепенно значение. Предвид обучаемостта на ИИ, се очаква резултатите да стават все по-добри; ако не се ограничи точно кое използване на ИИ е етично и допустимо в научната продукция, рискуваме пренасищане на научно-методологическата литература с вторично, генерирано съдържание.

Етично и отговорно използване на софтуер, позволяващ да се „очовечи“ текст, генериран от ИИ

Стремежът да се усъвършенства качеството, кохерентността и четимостта на генерирания текст с цел да се автоматизират редица действия, свързани с творческия процес на писане, води до паралелното разработване на софтуер, способен да „очовечи“ генерирания от ИИ текст, който може да звучи сухо, шаблонно, клиширано. Този вид софтуер на практика маскира факта, че първоначалната версия на текста е била генерирана от ИИ, като се открива поле за потенциални злоупотреби. По този начин се развива надпревара между компании, разработващи ИИ системи, базирани върху ГЕМ, които проверяват за наличие на ИИ-генерирано съдържание, и такива, които помагат да се скрие ИИ-генериран текст. Въпреки че едни от основните потребители на подобен софтуер изглежда са заети в областта на образованието и научната продукция, в декларацията за етично използване на Undetectable AI четем, например:

В компанията Undetectable.ai не одобряваме нито поощряваме използването на нашия продукт „за нарушаване на академичната почтеност или измама. Създали сме платформата си единствено с цел да се преобразува текст, създаден от ИИ, и насърчаваме всички потребители да го използват отговорно и съобразно етичните принципи. Undetectable не одобрява използването на тази технология за измама на системи за засичане на текст, генериран от ИИ в образованието“.

С технологията ни ние се ангажираме в борбата срещу нарушенията на академичната почтеност, като снабдяваме професионалисти от областта на маркетинга и творческите професии с ценни продукти. Ние вярваме, че отговорното използване на нашата платформа е от първостепенно значение за запазване на доверие и интегритет в индустрията[14].

Под формата на декларация компанията изглежда се дистанцира от потенциалните злоупотреби със софтуера, чиято основна маркетингова линия е, че обработеното с него съдържание успява да избегне контрола за наличие на ИИ-генериран текст от всички безплатни и комерсиално

[14] https://undetectable.ai/blog/ai-authorship-recognition-ai-detection-bypass/

достъпни системи за проверка[15]. Според Бизнес Инсайдър, целевата група за този софтуер са, в това число, създатели на съдържание, попадащо под защита на авторско право:

„Ценността на продуктите, предлагани от Undetectable.AI, се състои в това, генерираният от ИИ текст да изглежда създаден от човек и така да бъдат привлекателни за експерти в оптимизацията за търсачки, в рекламата и маркетинга, журналисти и други професионалисти, които биха искали да подобрят потока на съдържание и да поддържат качеството му на създадено от човек“. (Business Insider, May 8th, 2023)

Технологически гиганти от типа на Гугъл отбелязват следното по повод генерирано съдържание от ИИ в становище по въпроса:

– В рамките на системата за класиране на сайтовете и тяхното съдържание, която помага на търсачката да предлага определени резултати, екипът по качество на Гугъл поощрява качественото съдържание, без значение как е произведено – от човек или от ИИ;

– Принципите, според които се оценява съдържанието, са E-E-A-T: expertise, experience, authoritativeness, trustworthiness (професионално знание, опит, авторитетност, надежност), независимо от производителя, тъй като спам съдържание може да се произведе и от човек;

– Относно автоматично генерирано съдържание, политиката на Гугъл е следната: „Използването на автоматизация – включително с помощта на ИИ – за генериране на съдържание с основна цел да се манипулира рейтинговата система на резултатите от търсачката представлява нарушение на правилата за спам съдържание на компанията“[16]. Те уточняват, че не всяко автоматично генерирано съдържание е спам, понеже има множество ползотворни приложения на автоматично генерирано съдържание, като генериране на спортни резултати, на метеорологични прогнози и транскрибиране на реч.

Очевидно, линията, която очертават в Гугъл, засяга действия, които целят да се измами рейтинговата система за получаване на облаги. На преден план излиза идеята за ползите на ИИ при рационализиране на някои повтарящи се действия и за нарушаване на етичното използване, ако е налице манипулативна цел или умисъл за злоупотреба.

Представи за етично използване на генеративен ИИ

През последните година и половина в различните платформи за споделяне се увеличават видеоклиповете за помощ при използване на ИИ, като на преден план излиза темата за това как да се скрие, че текст е генериран от ИИ. Въпросът кое използване на ИИ е допустимо/етично в дадена област на човешката дейност, се поставя с нова острота. Прегледът, който следва, е направен върху една от най-използваните платформи за споделяне, които поддържат тежки файлове (съдържащи от няколко минути до няколко часа съдържание, позволяват онлайн стрийминг и т.н.) – YouTube. Изборът на платформата за видео споделяне се налага и поради предпочитането на видео-формата от потребителите (720 хиляди часа видео съдържание се споделя всеки ден - tubefilter.com).

Изборът на съдържание, което има шанс да бъде споделено, коментирано, харесано и т.н. – с други думи, съдържание, което ще бъде препоръчано от алгоритмите на платформата на голям брой потребители, зависи от горещите теми на деня. В месеците след пускането в свободен достъп на ChatGPT през ноември 2022 платформата се пълни с видео съвети, целящи да се разясни как да се използва технологията за оптимизиране на някои видове дейности (в частност, свързани със създаване на текст/съдържание), а от средата на 2023 година започват лавинообразно да се качват и споделят съвети с начини за избягване на детекция на ИИ-генериран текст. Тази тенденция логично отговаря на паралелно протичащия процес на разработване и усъвършенстване на софтуер, който прави проверка за наличие на ИИ-генериран текст. За необходимостта да се прави такава проверка първи алармират университетски преподаватели, които сигнализират за увели-

[15] На страницата на https://chatgptdetectors.com/ може да се достъпят основните продукти за проверка на ИИ-генерирано съдържание, които са основно насочени към обучителите и образователните институции.

[16] https://developers.google.com/search/blog/2023/02/google-search-and-ai-content

чаващи се случаи на студентски писмени работи, писани с помощта на ChatGPT (темата е достатъчно интересна, за да бъде отразена в медиите, насочени към широката общественост[17]). Фактът, че се търсят съвети за скриване на текст, генериран от ИИ, възможно сочи именно това, от което се опасяват редица учени: ползвателите на ИИ системите предпочитат лесното пред правилното (или етичното).

Таблица 1 съдържа преглед на основните теми с брой прегледи в платформата (от първите 3 страници резултати)[18]:

Таблица 1. Популярни теми, свързани с ИИ-базиран софтуер

Тема	Брой клипове	Общ брой прегледи
Using AI tools/ ChatGPT course	6	11.24 M
Bypass TurnItIn	21	1.7 M
Avoid AI writing detection	45	4.3 M
Writing with ChatGPT without getting caught	38	4.8 M
Remove plagiarism	17	3.6 M
AI tools for researchers	21	4.7 M

Тук трябва да добавим и материалите, които излизат с ключова дума „ethically“ (how to write with ChatGPT ethically), които не са толкова много, колкото представените в таблицата, и не привличат същия брой преглеждания:

Таблица 2. Ключова дума: етично използване на ChatGPT

Заглавие	Брой прегледи
10 ways to use ChatGPT to write research papers (ETHICALLY)	595 K
Using ChatGPT Ethically in Academic Writing: Best Practices and Guidelines	6.8 K
How To Automate Your Literature Review ETHICALLY Using ChatGPT (Prof. David Stuckler)	216 K
5 ethical ways to use ChatGPT for grant writing	5.8 K
How I ETHICALLY use ChatGPT to do my 9 months research in 3: From preparation to dissertation PART 1	4.1 K

Очевидно, интересът на пишещите в академичните среди към възможностите на ИИ е изключително силен, като етичната страна на въпроса не изглежда да се намира в центъра на вниманието.

Голямата адаптивност на учения, който вече 25 години работи в дигитална среда, който от повече от десетилетие се осланя на Гугъл продукти, за да си набавя научно-методическата литература (Carpenter 2012, Housewright et al. 2013), е в основата на възприемчивостта към новите технологии, въпреки въпроси от етичен и/или практически характер, които могат да възникнат по повод тяхната употреба (Vassileva & Chankova 2020). Възрастта на учения не е стабилен индикатор за одобрение/ възприемане/ използване на нови технологии, по-възрастните учени често са сред по-ентусиазираните ползватели на нови технологии и изследват с по-голям интерес новите възможности (Procter et al. 2010). Почти петнадесет години по-късно, когато разликите в техноло-

[17] AI breakthrough ChatGPT raises alarm over student cheating (ft.com) Universities warn against using ChatGPT for assignments (bbc.com)

[18] Систематичен преглед е безсмислен поради ежедневно увеличаващия се брой материали с подобна тематика. Броят прегледи е приблизителен (закръглен до 1000), към 26 юли 2024 г.

гическите умения на представителите на различните поколения са практически несъществуващи, голяма част от процесите, съпътстващи учения (публикационна активност, разпространение на резултатите от научните изследвания и др.), се случват в дигиталната среда. С разработването на споделената мрежа стоят ред проблеми, касаещи етиката на научноизследователския процес: въпросът доколко представителна картина на научната област дават резултатите от дигитални търсачки (Weller 2011); опасенията, че част от цитираните статии не са реално прочетени (или е прочетена само анотацията – Rowlands et al. 2008; Kroll & Forsman 2010); като мотивацията на учените да използват новите технологии изглежда е доколко оценяват ползата им и лекотата на използването им (Seyal et al. 2002). В България изследване показа известна резервираност на учените от хуманитаристиката спрямо новите технологии, но готовност да ги приемат при определени условия, като новите технологии не изместват старите, а ги допълват (Vassileva & Chankova 2020).

Основният акцент във видеоматериалите е върху рационализирането на процеса на изследванията: колко по-бързо и по-лесно се извършват някои дейности (търсене и обработване на научно-методическа литература, извличане на основни/приносни моменти от литературата, получаване на отговор на въпрос на база на научно-методическата литература, генериране на текст и проверка за съвпадение/плагиатство). Въпросът, който не се обсъжда в тези материали, е следният: ако ИИ събира корпус от статии за литературен обзор по зададени параметри, друг ИИ извършва обобщение на теми/приноси/заключения по зададени параметри, трети ИИ формулира въпрос/тема за научното изследване, какво и колко чете самият учен, който оперира тези ИИ технологии? Самите автори на материалите често сравняват възможностите на новите технологии с личния си опит – сбор на материали, обработване, обобщаване, извличане на съществено и т.н., който е правен без специален софтуер, трудоемка дейност, която изисква цял набор от специални компетентности от страна на учения – и лекотата, с която подобни дейности (или части от дейности) могат да бъдат извършени от технологии с интегриран ИИ. Как това би се отразило върху формирането на тези компетентности? Или необходимостта от подобни компетентности ще отпадне? Да уточним, че най-застрашени са така наречените компетентности за обработване на информация:

– Разпознаване на необходимостта от информация;
– Умение за достъпване на информация;
– Умение за оценка на информацията;
– Умение за обобщаване и извличане на съществено от информацията;
– Умение за комуникиране на информацията (по Durbin 2009).

Формирането на някои други компетентности също така са под заплаха от генеративния ИИ, като критично и творческо мислене, аналитични умения, умения за интерпретиране и абстрактно мислене (Datoo & Siddiqui 2024), което в крайна сметка може да има сериозни последствия за формирането на следващите поколения, преминаващи през образователна система, доминирана от генеративен ИИ. В този смисъл, някои формулировки, които се използват в дискусията, са доста неопределени: ИИ следва да бъде не средство за писане, а „помощник“ при процеса на писане, като човекът трябва да запази своята „автентичност“. Подобни формулировки в контекста на обучителни видеоматериали на тема как да научим ChatGPT да пише с нашия стил изглеждат най-малко неискрени (ако не и откровено цинични).

Един от въпросите, които повдигат някои от авторите на подобни публикации, е за процента фалшиви положителни резултати, които могат да бъдат получени при проверка за наличие на ИИ-генериран текст: т.е. когато софтуерът определя даден текст за генериран от ИИ, а текстът всъщност е писан от човек. Тази възможност демонстрира още веднъж слабостите на ИИ-базирания софтуер, но и посочва слабото звено на система, в която, въпреки голямото разнообразие на възможните приложения, първично остава утилитарното пестене на работа от страна на различните заинтересовани в образователната система.

Още един въпрос, който си задават част от споделящите видеоматериали на тема ИИ, е дали не е достигнат апогея на възможностите на този вид ИИ-базиран софтуер. Поводът за подобни въпроси са изследвания, които стабилно корелират честотата на появяване на дадено понятие в базата данни, залегнала в ГЕМ, както и приликата с базата данни, с резултатите и производител-

ността на моделите (напр. Udandarao et al. 2024, Mayilvahanan et al. 2023). На базата на подобни изследвания се прави предположението, че независимо колко се увеличава базата данни, постепенно индикаторът на производителността на моделите ще се стабилизира, вместо да нараства, което на практика ще означава, че е достигнат максимумът на възможностите на модела[19]. Имайки предвид, че моделите не разполагат със семантично разбиране (Titus 2024), това означава, че независимо от големината на модела, той най-вероятно не би могъл да отговори правилно на въпроси, подобни на следния въпрос на Pinker:

„Когато попитах ChatGPT „Ако Мейбъл е била жива в 9 ч. сутринта и в 5 ч. следобед, била ли е жива на обяд?“, той ми отговори „не е уточнено дали Мейбъл е била жива на обяд. Знае се, че е била жива в 9 и в 5, но няма информация дали е била жива на обяд.“ Значи, той не разбира основни факти за света, като това, че хората живеят в продължителни отрязъци от време и че когато човек умре, той остава мъртъв - именно защото не е срещнал текст, в който това буквално е споменато.“ (Pinker 2023)

Етичните въпроси около използването на ИИ в академичната област засягат прозрачността (какво и как точно е използвано), отговорността (авторът отговаря за достоверността и истинността на предложения текст), липсата на намерение да се получат незаслужени облаги (от страна на студента - да получи оценка за работа, която не е писана от него, от страна на учения – да получи признание за принос, който не е негов[20]). Засилващите се темпове, с които се променя цялата академична екосистема под въздействие на новите технологии, е обективна пречка да се направят прогнози за това в каква насока по-точно ще се развие приложението на генеративния ИИ. Поради лекотата на използване и огромния потенциал за рационализиране на някои видове дейности, възможно е в един начален етап да се наблюдава лавинообразно увеличаване на използването му – частично към този извод сочат резултатите от процесите на споделяне на съдържание в YouTube. Възможно е след първоначален интерес да дойде спад на актуалността на тези технологии: подобен процес вече се наблюдава на комерсиалните пазари на разработчици на ИИ, където след високо търсене и огромен интерес от страна на инвеститори се наблюдава спад, особено след някои скандали за измами[21] и провали[22].

Заключение

Конкретното въздействие на генеративния ИИ се оценява в загуба на работни места в някои специализирани области, общ принос към увеличаване на БВП, висок потенциал за автоматизиране на някои дейности, свързани с производство на специализирано знание и/или решение (между 40 и 87%), като същевременно се увеличава процентът хора, които използват генеративен ИИ за работа или извън нея (по данни на McKinsey от 2023 г.[23]). От друга страна, ситуацията на безконтролна конкуренция между разработчици, които могат да не следват принципите на етично разработване, както и приложението на подобни системи в някои чувствителни сфери на човешката дейност, потенциално носят значителен риск за човешкото общество (напр. Hinton, наречен Кръстникът на системите на генеративен ИИ[24], или още Gawdat 2021). Сред ползвателите преобладава утилитарен, прагматичен подход към тази технология, който се обуславя в краткосрочна перспектива от ползите, които те се стремят да получат. Дългосрочната перспектива, която би следвало се осланя на морално-етични принципи за ненанасяне на вреда, не изглежда да влияе при вземането на решение дали да се прибегне или не до тази технология за определена дейност.

---

[19] Например Has Generative AI Already Peaked? https://www.youtube.com/watch?v=dDUC-LqVrPU.

[20] Това определение на практика покрива широкото понятие за плагиатство, при което се търсят заслуги за извършено от други хора.

[21] https://www.sec.gov/newsroom/press-releases/2024-36

[22] https://tech.co/news/list-ai-failures-mistakes-errors

[23] https://www.mckinsey.com/~/media/mckinsey/featured%20insights/mckinsey%20explainers/whats%20the%20future%20of%20generative%20ai%20an%20early%20view%20in%2015%20charts/whats-the-future-of-generative-ai-an-early-view-in-15-charts.pdf

[24] https://www.youtube.com/watch?v=qrvK_KuIeJk

Така в името на автоматизирането на определен вид дейности може да се пренебрегнат ползите на същите тези дейности за формиране на ключови компетентности, които имат широко приложение.

Освен това, предвид множеството сфери на приложение на генеративния ИИ, както и многобройните действащи лица, които потенциално биха имали различни (възможно противоречащи си) цели и намерения, прави невъзможна оценката (както и прогнозата) на реалното въздействие в конкретна област. Има една идея, около която изглежда се обединяват голяма част от заинтересованите лица – разработчици, издатели, учени, създатели на научна политика, законотворци – необходимостта да се регулира използването на технологията, въпреки че всички признават за съществуването на времева пролука, необходима за осъзнаването на потенциалните ползи и рискове, което предвид развиващата се област затруднява още повече оценката на реалното въздействие на технологията, от една страна, и адекватността на приетите мерки и политики, от друга.

Ето защо в момента е налице спешна необходимост от създаване на насоки/правила за използване на системите с ИИ в съответните области. В контекста на тази статия бихме препоръчали МОН, съвместно със специалисти от различни области на науката и образованието, да излезе с такива насоки, които да могат да се приемат от образователните и научни институции и издатели, като се даде възможност те да се модифицират в съответствие с конкретния контекст (например хуманитарни, обществени науки, точни науки и т.н.). Както по-горе беше изяснено, съществуват вече достатъчно документи на ниво ЕС, както и редица примери от университети в чужбина, които могат и трябва да послужат като ориентир. Правилата би трябвало да останат гъвкави, с възможност да се актуализират в зависимост от посоката на развитие на технологиите с ИИ. Тези правила могат да се включат към вече съществуващите Етични кодекси, или да бъдат отделни документи.

Предприемането на горепосочените мерки ще доведат до въвеждането на ред в хаоса, който в момента съществува в България както по отношение на използването на тези технологии в образованието (особено във висшето), така и в научните изследвания и тяхното публикуване. Разбира се, всичко това би трябвало да се подкрепи с разяснителна медийна кампания, целяща широката публика.

ЛИТЕРАТУРА/ REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). „On the dangers of stochastic parrots: Can language models be too big?“– In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* **(pp. 610-623).**

Carpenter, J. (2012). Researchers of Tomorrow: The research behaviour of Generation Y doctoral students. *Information services & use***,** *32*(1-2), 3-17.

Datoo, A. K., & Siddiqui, K. A. (2024). ChatGPT and Death of an Author. *Critical Humanities***,** *2*(2), 5.

Dwivedi, Y. K. et al. (2023). Opinion Paper: „So what if ChatGPT wrote it?“ Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management***,** *71***, 102642.**

EduKitchen (2023, January 20). Chomsky on ChatGPT, Russia, Education and the unvaccinated. [video] https://www.youtube.com/watch?v=IgxzcOugvEI.

Foltýnek, T. (2024). „Deep dive into generative AI and large language models“. Presentation at WCRI 8th World Conference on Research Integrity 2-5 June 2024, Athens, Greece.

Frye, B. L. and Chat GPT (2023). Should Using an AI Text Generator to Produce Academic Writing Be Plagiarism?, *33 Fordham Intell. Prop. Media & Ent. L.J.* 946.

Gao, C. A. et al. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine***,** *6***(1), 75.**

Gao, Y., Wang, R., Hou, F. (2023). How to Design Translation Prompts for ChatGPT: An Empirical Study. *arXiv preprint arXiv:2304.02182v2*.

Ganjavi, C. et al. (2023). Bibliometric Analysis of Publisher and Journal Instructions to Authors on Generative-AI in Academic and Scientific Publishing. [Accessed 20.06.2024] *arXiv*, https://doi.org/10.48550/arXiv.2307.11918.

Gawdat, M. (2021). *Scary Smart: The Future of Artificial Intelligence and How You Can Save Our World.* London: Pan Macmillan.

Hendy et al. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv:2302.09210*.

Housewright, R., Schonfeld, R. C., & Wulfson, K. (2013). Ithaka S+ R/Jisc/RLUK UK Survey of Academics 2012. *London: University of London*.

Kroll, S., & Forsman, R. (2010). *A slice of research life: information support for research in the United States*. Available at https://core.ac.uk/download/pdf/30682718.pdf. [Accessed 20.06.2024].

Marcus, G. & Davis, E. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust***. Vintage Books Edition.**

Peng, K. et al (2023). Towards Making the Most of ChatGPT for Machine Translation. *arXiv preprint arXiv:2303.13780v1*.

Perera, P., & Lankathilaka, M. (2023). AI in Higher Education: A Literature Review of ChatGPT and Guidelines for Responsible Implementation. *International Journal of Research and Innovation in Social Science*. doi:10.47772/ijriss.2023.7623.

Pinker, S. (2023). Will ChatGPT supplant us as writers, thinkers? *The Harvard Gazette*. February 14, 2023.

Procter, R. et al. (2010). Adoption and use of Web 2.0 in scholarly communications. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4039-4056.

Seyal, A. H., Rahman, M. N. A., & Rahim, M. M. (2002). Determinants of academic use of the Internet: a structural equation model. *Behaviour & Information Technology*, *21*(1), 71-86.

Titus, L. M. (2024). Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, *83*, 101174.

Van Dis, E. A. et al. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224-226.

Vassileva, I. (2024). Machine versus Human Translation – High-Resource versus Low-Resource Languages. A Case Study. (под печат).

Vassileva, I. & Chankova, M. (2020). Scholars' information exploitation habits in multimedia environment. In: Vassileva, Chankova, Breuer & Schneider (eds.), *Digital Scholar: Academic Communication in Multimedia Environment*, pp. 63-92. Berlin: Frank & Timme.

Weller, M. (2011). *The digital scholar: How technology is transforming scholarly practice* (p. 208). Bloomsbury Academic.

Куманова, Е., Даскалова, С. (2024). Правен режим на използването на изкуствения интелект в образователния процес. https://www.conf-dte.bg/docs/2024/p-81.pdf. [Accessed 30.06.2024] // Kumanova, E., Daskalova, S. (2024). Praven rezhim na izpolzvaneto na izkustweniya intelekt v obrazovatelniya prozes.

Източник на финансиране: Изказваме благодарност на ФНИ, който финансира настоящото изследване в рамките на проект: „Сериозността на академичното плагиатство в нагласите на учени, студенти и създатели на научна политика в България„ по договор с ФНИ № КП-06-Н70/9 от 13.12.2022 г.