

**Медийна и дигитална комуникация
Media and Digital Communication**

**Дезинформацията в ерата на изкуствения интелект:
предизвикателства пред дефинирането
и разпознаването**

DOI 10.5526/WGRM2478

Тихомира Дончева

Софийски университет „Св. Климент Охридски“

Имейл: tdoncheva@phls.uni-sofia.bg

Абстракт: Разпространението на дезинформацията е сред основните заплахи за демократичните процеси, а бързо развиващото се поле на изкуствения интелект (ИИ) допълнително усложнява противодействието към нея. Европейският съюз (ЕС) е сред пионерите в опитите за регулация на ИИ, а от август 2024 г. следва и строгите правила. Какво обаче означават термините „дезинформация“ и „изкуствен интелект“? Как тези термини се пресичат и какви са реалните опасности (и възможности), за които предразполага синхронизацията на двете явления? Това са двата изследователски въпроса, които се поставят в статията. Целта е да се проследят и систематизират основните тенденции при дефинирането на двете понятия, като се поставят в контекста на регулиращите цифровите услуги правила в ЕС: Законодателният акт за цифровите услуги и Законодателният акт за изкуствения интелект. Хипотезата е, че разбирането на двете понятия от широката публика е трудно, заради широкообхватността и труднодоказуемостта на терминологията. Изследването се осъществява при прилагане на методите кабинетно проучване, анализ и синтез. В дискусийната част на статията са разглеждани някои от основните опасности и възможностите, които се откриват при синхронизацията на двете явления.

Ключови думи: дезинформация, изкуствен интелект, термини, Европейски съюз (ЕС), регулация, Законодателен акт за цифровите услуги (DSA), Законодателен акт за изкуствения интелект (AI Act).

Disinformation in the Age of Artificial Intelligence: Challenges in Definition and Recognition

Tihomira Doncheva

Sofia University “St. Kliment Ohridski”

E-mail: tdoncheva@phls.uni-sofia.bg

Abstract: Disinformation is one of the greatest threats to democratic processes, and the rapid development of artificial intelligence (AI) further complicates the fight against its spread. The European Union (EU) has taken a pioneering role in regulating AI, introducing the strictest rules since August 2024. But what do the terms “disinformation” and “artificial intelligence” really mean? How do they interact, and what are the potential dangers (and opportunities) that arise from their intersection? This article explores these questions by examining the key trends in an attempt to define these concepts, while situating them within the context of EU regulations such as the Digital Services Act (DSA) and the Artificial Intelligence Act (AI Act). The hypothesis is that the general public finds it difficult to fully understand these concepts due to their broad scope and the challenges in defining their terminology. Through desk research, analysis, and synthesis, this paper assesses whether the frequent use of these buzzwords means they are largely understood or misunderstood. In the discussion section, the article summarises the main challenges and opportunities, posed by the synchronisation of these two phenomena.

Keywords: disinformation, artificial intelligence, terminology, European Union (EU), regulation, Digital Services Act (DSA), Artificial Intelligence Act (AI Act).

Увод

Oxford Dictionary определя постистината (post-truth) за дума на годината през 2016 г. Контекстът е свързан с политическата обстановка около избирането на Доналд Тръмп за президент на САЩ и с вота за Брекзит във Великобритания [1] – свят, в който хората все по-често се доверяват на чувства и убеждения отколкото на факти. [2] През същата година Politifact обявява фалшивите новини (fake news) за лъжа на годината (класация, в която организацията определя най-абсурдните лъжи в американските медии), повлияна от „символизма в политическа година, изпълнена с необузвано възмутително лъгане“ и невъзможността да избере само една лъжа. [3] Така тези два термина, свързани с различаването на истинска информация от неистинската, навлизат в общественото ползрение, за да се превърнат в една от най-горещите теми за демократичния свят през 21. век. [4] [5]

Много преди 2016 г. обаче въпросът какво е истина и факт в информационното съдържание вълнува експерти, научни представители, между-

народни организации и правителства. [6] Тези явления остават някак встра-ни от вниманието на широката публика... до 2018 г., когато избухва скан-далът с Cambridge Analytica и се повдига завесата за необятните възмож-ности, които технологичният напредък отваря за разпространението на ин-формация – и вярна, и невярна. [7]

Оттогава насам словосъчетанията фалшиви новини и постистина са избягвани [8] и думите „дезинформация“ (disinformation) и „мисинформа-ция“ (misinformation) (отново термини с дълбока история още от Римската империя [9]) навлизат все повече в публичното пространство (а в акаде-мичните среди водят до експлозия от съдържание). Тази лавина от публична и експертна оценка по темата води до общо объркване и неразбиране на основната терминология и допълнително допринася за разпространението и влиянието на невярна информация.

В края на 2022 г. още един технически пробив допълнително услож-нява вече нажежения климат около дезинформацията: Open AI отваря за свободен достъп генеративния чатбот ChatGPT [10], което превръща „из-куствения интелект“ (ИИ) в неизменна част от ежедневието на обикновения гражданин и дава началото на ерата на широкодостъпния ИИ. Създаването на невярна информация вече се случва за секунди, което позволява достъп до по-голяма аудитория, на различни езици, едновременно. Какво обаче се има предвид под „изкуствен интелект“, „чатбот“ и „голям езиков модел“ – словосъчетания, с които често се описват популярни технологични средства като ChatGPT? Осмислянето на терминологията е трудоемко, а използ-ването ѝ в контекста на разпространението на дезинформацията – широко разпространено.

Всички тези събития през последното десетилетие доказват значи-мостта на борбата с дезинформацията в ерата на ИИ за публичната дис-кусия. [11] Изброените ключови примери показват и колко е важно явле-нието да се разглежда от няколко различни точки – правно-регулаторна, журналистическа, но и академична. Загатва се обаче за една тревожна тен-денция – затова как дадени ключови думи се използват прекомерно в пуб-личния дискурс и как това може да доведе до изпразване на тяхното зна-чение.

Въпросът е дали термините „дезинформация“ в контекста на „изкуст-вения интелект“ не са тръгнали по пътя на постистината и фалшивите новини към изпразването от съдържание. В експертните среди тези две яв-ления яростно се избягват като лишени от смисъл думи, които само привли-чат внимание на широката публика и водят до допълнително поляризиране на комуникационния език.

В опит за принос към гореспоменатата дискусия, тази статия цели да обърне внимание, че термините „дезинформация“ и „изкуствен интелект“

поотделно са трудни за дефиниране. Прекомерното им използване в публичното пространство води до неразбирането им от широката публика. Следователно еволюцията на явлениято „разпространение на дезинформацията в ерата на ИИ“ трябва да се проследи и систематизира, като се поставят в контекста на регулиращите цифровите услуги правила в ЕС.

Основната хипотеза, която разглежда статията, е, че разбирането на двете понятия от широката публика е трудоемко, заради широкообхватността и труднодоказуемостта на терминологията. Настоящият труд ще се опита да заключи чрез общонаучните методи кабинетно проучване, анализ и синтез на наличната литература дали честото използване на тези думи не е опит за привличане на вниманието на аудиторията, без обаче да се взима предвид значението на думите. Зададени са два основни изследователски въпроса: Какво се има предвид под тези популярни термини (buzzwords) „дезинформация“ и „изкуствен интелект“? Как двата термина се пресичат и какви са реалните опасности (и възможности), за които предразполага синхронизацията на двете явления? Статията предлага още насоки какво и как да се разбира при използването на тези термини.

Теоретичен обзор

Да се направи подробен литературен обзор по темата е почти невъзможно, тъй като едни от първите сведения за дезинформация датират от Римската империя, а изкуственият интелект се разнищава от експерти в най-разнородни направления – от юристи до инженери. Затова в рамките на тази статия обзорът е лимитиран до разглеждане на двете твърдения, така както те си взаимодействат едно с друго.

Дезинформацията в ерата на цифровизацията и негативните ефекти от разпространението ѝ се отразяват по-бързо в обществото и достигат по-голям обхват, като се наблюдава увеличаване на нивото на невярна информация онлайн в световен мащаб. [12] През последните години ЕС е сред пионерите, които се опитват да урегулират разпространението на дезинформация [13], приемайки ключовите законодателни актове за цифровите услуги и за цифровите пазар. [14]

Но още по-важно – разпространението на дезинформация пресича развитието на ИИ, затвърждавайки мястото си сред най-големите трудности за социума през 21. век. ЕС – отново пионер – не закъснява и приема още един законодателен акт – този за изкуствения интелект, като не пропуска да направи връзка между дезинформацията и изкуствения интелект в текста. Европейските избори през 2024 г. са пример за необятността от възможности, които ИИ разкрива пред създателите и разпространителите на дезинформация.

В началото на 2025 г. едно видео обръщение на Марк Зукърбърг, създател и директор на Мета, дава нов тласък на дискусиите за разпространение на дезинформацията в ерата на ИИ – Мета спира проверката на факти от трети лица в своите социални платформи на територията на САЩ, но засега оставя модерацията на съдържание в Европа. [15] [16] Това решение на Мета със сигурност ще доведе до директен сблъсък на компанията с политиките на ЕС за засилване отговорността на онлайн платформите. [17] [18] [19]

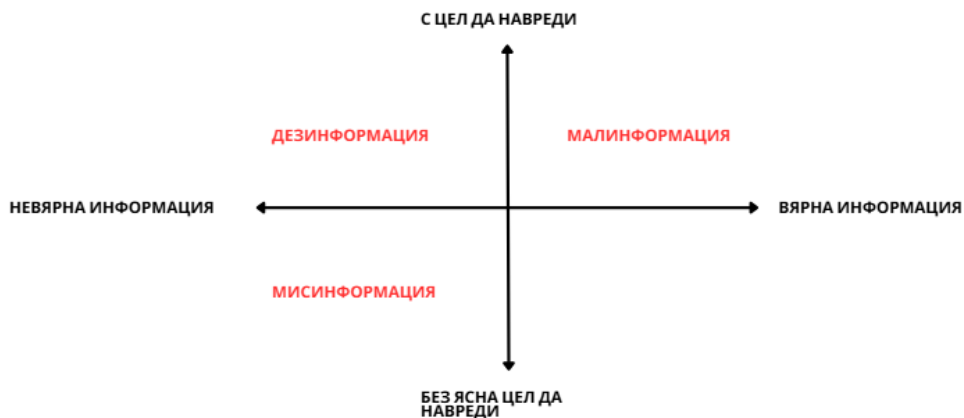
Методика

Използваната методология включва общоприети качествени методи като кабинетно проучване на научна литература, официални документи и институционални източници и други, сред които медийни публикации; анализ и синтез на наличната информация. В настоящия труд няма емпирични изследвания и резултати, тъй като целта е да се разгледат и систематизират съществуващи дефиниции. Приносите на статията са представени в дискуссионната част, където са изведени предложения за употребата и значението на двата термина в контекста на неекспертни и научни разговори.

Дезинформация, мисинформация и малинформация: терминологичен обзор

В опит да се представи дефиниция за дезинформацията, тази статия използва даденото от UNESCO обяснение за информационното разстройство. [20] Според организацията, дезинформацията, мисинформацията и малинформацията са три свързани явления, които са разположени на спектъра неистина-истина. Ако начертаем диаграма с квадранти, различните проявления на информационното разстройство ще се подредят по следния начин: невярна информация с цел да навреди (дезинформация), невярна информация без ясна цел да навреди (мисинформация) и вярна информация с цел да навреди (малинформация). Дезинформацията се отнася за изфабрикувано, нарочно манипулирано съдържание, което е създадено с цел да навреди. Мисинформацията се разполага вляво на дезинформацията, като отново съдържанието е невярно (включително неволна грешка, грешна връзка между твърдения и подвеждащо съдържание), което обаче не се разпространява с ясната цел да навреди. Малинформацията, от друга страна, е разположена вдясно на спектъра и се отнася за информацията, която в същността си е вярна, но се разпространява с ясната цел да навреди (репутационно на някого, публикуване на лична информация, която не е в интерес на обществото, или става въпрос за нарочно подменен контекст на информацията, дата на публикуване или друго отнасящо се за вярна информация). Информационното разстройство предлага работна рамка за категоризиране

на информацията, която показва колко е трудно да се дефинира дезинформацията. Как се доказва ясната цел за вреда, което различава дез- от мис-информацията (фиг. 1)?



Фиг. 1. Диаграма на информационното разстройство

В българския контекст пионер в превеждането и адаптирането на термините е Асоциацията на европейските журналисти (АЕЖ-България), която публикува пътеводителя за проверка на фактите, в който дефинира мис-, дез- и малинформация (базирано на дефинициите, дадени от First Draft). [21] В документа дезинформацията се дефинира като „изфабрикувано или умишлено манипулирано аудио или видео съдържание, умишлено разпространявани слухове и конспиративни теории“. [22] В сравнение с дефиницията на UNESCO, тук отново се засяга темата за невярна информация (изфабрикувано или манипулирано съдържание), както и че се отнася за умишлено манипулиране и/или умишлено разпространение, т.е. отново се набляга на умисъл, цел за вреда. Макар авторитети от First Draft/ АЕЖ-България да лимитират дезинформацията до аудио и видео съдържание, дезинформативното съобщение може да е всичко – както текст (например статия, доклад, документ), аудио и видео съдържание, така и друг тип информационно съобщение (например таблица, графика, снимка и др.). От друга страна, мисинформацията и според авторите на пътеводителя се отнася за невярна, но неволно разпространена информация: „неволни грешки, като неточни заглавия на снимки, дати, объркани статистически данни или превод, позоваване на сатирични сайтове като реални информационни източници“. [23] Малинформацията и тук се определя като вярна, но разпространена с дадена цел информация: „Умишлено публикуване на лична информация в името на личен или корпоративен интерес. Умишлено изменение

на контекста, датата или времето на публикуване на автентично съдържание“. [24]

Въпросът за умишленото разпространение възниква и в тази дефиниция. Това подразделение на информационното разстройство – на дез-, мис- и мал- (както го наричат от First Draft), публикувано в пътеводителя за проверка на факти, има за цел да подпомогне проверителите на факти (factcheckers) да разследват информационния поток. В този случай много по-лесно може да се докаже дали дадено невярно съдържание е разпространено с цел да повлияе негативно, или по-скоро става въпрос за неволна грешка (основната разлика между мис- и дезинформация). Много често проверителите на факти успяват да докажат, че дадено съобщение е с дезинформационен характер, тъй като е част от по-голямо таргетирано манипулиране на информационната среда (например чуждестранни информационни манипулации и намеси (foreign information manipulations and interference [25]), хибридни войни [26] и др.).

Макар да се разглеждат само два източника за дефиниция на дезинформацията, става ясно колко е трудно да се даде кратко, точно и ясно определение за явлението, без допълнително да се появяват въпроси за доказуемост на нагласите на източника на информацията или без да се добавят нови термини, които допълнително затрудняват аудиторията.

Затова е важно, когато дадена информация се представя за дезинформация, то контекстът да е свързан с правно-регулаторна рамка, каквато например очертава ЕС. Подобна на горните дефиниции е и дефиницията за дезинформация и мисинформация, дадена от ЕС. Дезинформацията според ЕС е „невярно или подвеждащо съдържание, което се разпространява с намерение да заблуди или осигури икономическа или политическа печалба и което може да причини обществена вреда“, докато мисинформация е „невярно или подвеждащо съдържание, което се разпространява без намерение за вреда, но което може да доведе до вреда“. [27] Разликата между дефинициите, дадени по-горе и тук, е в дефиниращата организация. Дефинициите на UNESCO и на First Draft са насочени най-вече към журналисти, медийни експерти, академичната общност, но и към обикновената аудитория. Дефинициите, дадени от ЕС, също би следвало да се използват от всички, но с тази разлика, че ЕС приема законова рамка, с която може да зададе ясни критерии, по които да се определя дали дадено невярно съдържание е разпространено с цел да навреди или не, и впоследствие да глоби създателите и разпространителите на дезинформация.

Мария Юркова представя изчерпателен исторически преглед и синтез на развитието на европейското мислене по темата за дезинформацията в своята докторска дисертация. [28] Накратко – ЕС е приел мултидисциплинарен подход, който включва регулация Законодателния акт за цифровите

услуги (DSA) и Законодателния акт за цифровите пазари (DMA), отговорност от институции и индустрия, проверка на факти и стимулиране на добрата журналистическа практика. Макар и DSA [29] да не е регулация само и единствено на дезинформацията, голяма част от правилното функциониране на цифровия пазар (целта на акта) е тясно свързана с противодействието на дезинформацията. Липсва обаче определение на дезинформацията, което да е включено в регулаторния документ, макар от Европейската комисия да твърдят, че DSA има за цел митигация на системни рискове, сред които дезинформацията. [30] Два са начините, по които дадено дезинформационно съдържание може да се обори според DSA: ако съдържанието е незаконно или ако заплашва за обществена вреда. В първия случай доказано дезинформационно съдържание трябва да се премахне от цифровите посредници (включително социални мрежи), ако се счита за незаконно според законовата рамка на страна членка. Тук се включва и задължението на онлайн платформите да осигурят механизми, по които всеки може да сигнализира дадено незаконно съдържание (включително каквото лицето смята за дезинформационно съдържание). Въпросите отново са повече от отговорите в контекста на разбирането кое съдържание е незаконно и кое – законно, дискусия, която изисква подробно познаване на законите на страните членки. От друга страна, DSA напълно забранява таргетираните реклами на базата на лична информация, както и задължава големите платформи (с над 45 млн. потребители) да предотвратят разпространението на вредна информация (вкл. дезинформация) без задължително съдържанието да е незаконно. Цифровите медиатори с над 45 млн. потребители месечно са задължени да следят за и да адресират системни рискове (вкл. в контекста на алгоритмите за реклама). [31] В пакета с мерки за дигитална регулация на ЕС влиза и DMA [32], но той от своя страна се отнася за достъпни, честни и справедливи цифрови пазари и не е директно свързан с дезинформацията.

В обобщение до тук: за дезинформацията се говори навсякъде – от академичните и експертни кръгове до регулаторните органи. Макар навсякъде дефинициите да си приличат (разликата между дез- и мисинформация е в целта на разпространението), терминологията не е общодостъпна и лесна за разбиране. Регулациите на ЕС – макар и сред най-добрите в световен мащаб по темата – оставят много отворени врати за интерпретация и не са толкова лесноразбираеми от широката аудитория (макар и тя да е включена като участник в мерките за ограничаване на дезинформацията). Ако дадено лице не е експерт по темата, не е обучен в проверката на факти, или не е запознат тясно с тематиката, много лесно може да оприличи дадено съдържание за дезинформационно, без да може да защити умишления характер, който различава дез- от мисинформация. Резултатът в ежедневието е информационна среда, в която дезинформацията се среща често в речта на

гражданите, приема се за сериозен риск за демокрацията (86% средно за ЕС) и се радва на сравнително висок процент (71% средно за ЕС) за идентифициране на невярно съдържание (на базата на самооценка). [33] Но макар да е толкова често припознавана като проблем, дезинформацията е все така често срещана, а самата дума много скоро може да загуби смисъла си от тази прекомерна употреба (подобна на постистина и фалшиви новини).

Изкуствен интелект: терминологичен обзор

Изкуственият интелект в най-общия смисъл е интелигентна компютърна система, която изпълнява задачи, изискващи човешка интелигентност. [34] Самото понятие възниква още през 1955 г. от Джон Маккарти, който го описва като наука и инженерство на създаване на интелигентни машини. [35] През 2007 г. Маккарти публикува обновена статия под формата на въпроси и отговори, които дават основополагащи дефиниции за полето, сред които разделяне на подвидове ИИ и основни приложения. [36] Съществуват две главни категории дефиниции за ИИ. През призмата на абстрактното, ИИ се сравнява с човешките способности за извършване на дадена задача. Именно в тази категория ИИ се представя за софтуер, програма или система, действащ аналогично на човешка реакция. Много често обаче дадени системи се наричат „интелигентни машини“ или „интелигентен софтуер“, но всъщност наименованията са неточни и неотговарящи напълно на определението за аналогичност с човешкото поведение. Втората категория дефиниции разглежда ИИ като способен да изпълнява задачи, поставени чрез алгоритъм, но и на базата на неговата самостоятелност и възможност за самообучение. [37] Широкоприетото разделение на ИИ включва: тясноспециализиран изкуствен интелект (решава определена задача), общ изкуствен интелект (универсални познания, може да решава разнородни човешки задачи, да мисли и да твори), изкуствен суперинтелект (по-добър от хората във всички задачи) и ИИ с човешко самосъзнание (цели, желания). През 2005 г. футуристът Рей Курцвел предрича, че в близкото бъдеще скоростта на технологичните промени ще бъде толкова бърза, а въздействието им толкова голямо, че човешкият живот ще бъде необратимо трансформиран. [38] Той нарича този период сингулярност, следващата стъпка след общия ИИ и тясноспециализирания изкуствен интелект.

Генеративният изкуствен интелект обаче е най-тясно свързан с темата за дезинформацията. Този вид системи се отнася за големи машинни модели, които са способни да създават ново съдържание (текст, картина, видео), след като вече са обучени на базата на големи бази данни от създадено от човека съдържание. Чрез въпроси и отговори човек може да общува с машината, да дискутира, да получава насоки, съвети и други. Най-

важното е, че генеративният ИИ разчита на вече научени модели и е събирателен термин за цялостната инфраструктура (вкл. математическия модел, обработката на данни, както и компонентите на потребителския интерфейс). С други думи, ИИ използва алгоритми и математически модели, които се обучават от бази с данни (data sets) чрез много примери. Това е т.нар. процес на машинното обучение. Тъй като генеративният ИИ разчита на алгоритми, дълбоките невронни мрежи (които подобно на човешкото съзнание могат да решават сложни проблеми) са добре приложими за генериране на съдържание, тъй като те могат да бъдат проектирани с помощта на различни архитектури за моделиране на типове данни като например последователни данни (какъвто е човешкият език) или пространствени данни (каквито са например изображенията). Тук основните машинни модели са GPT-4 (генерира текст по текстово задание), FLUX.1 (генерира картини по текстово задание) и SpeechT5 (генерира човешки говор по даден текст). Основните функции на този вид модели са да класифицират информация, да предсказват информация, да разпознават глас от аудио и видео, да разпознават текст от картина, да синтезират глас и видео от текст, да препоръчват релевантно съдържание, да отговарят на въпроси и др. [39]

Генеративните системи за ИИ, които използват големи езикови модели, не са новост и работят както с отворен достъп, така и със затворен. Интересът към този вид ИИ от широката публика експлодира през 2023 г., когато OpenAI отваря леснодостъпния модел за непрофесионалисти ChatGPT. В последните две години всички софтуерни гиганти интегрират помощник с ИИ в своите продукти: например Google използва няколко модела включително Gemini, Microsoft - Phi, Orca, както и ChatGPT. Китайският гигант Baidu разработва Ernie, а в началото на 2025 г. китайски учени представят на свободния пазар и DeepSeek – пряк конкурент на ChatGPT, но в пъти по евтин. [40] [41]

Един от основните проблеми на чатботовете с големи езикови модели се корени в ограниченията на знанията и признаците на халюцинации, които понякога показват. [42] Това е заложено в начина, по който моделите работят: тъй като машинното обучение залага на най-вероятните алгоритми, за да произведе отговор на задание, а не непременно на верния отговор. Халюцинациите при ИИ представляват именно отговор на задание, който изглежда правдоподобен като информация, но всъщност може да е пълен с грешки и невярна информация. [43] [44] При поглед малко по-нагоре в статията се вижда, че всъщност непълна и/или невярна информация, разпространявана без умисъл да навреди, е мисинформация – именно по този начин двете явления се преплитат и ИИ допълнително усложнява вече сложната обстановка на информационното разстройство. С други думи, възможна е ситуация, в която големите езикови модели произвеждат съдър-

жание, което не е базирано на точни факти, които са леснопроверими, а на собствените си предположения и предразсъдъци. [45]

Тъй като машинните модели се тренират върху големи набори от данни и се учат в разговорите си с човека, те имат определена дата – начало, от което знанията им започват. [46] Това представя още един проблем на ИИ: моделът използва само научената информация от тази начална дата. В допълнение ИИ може да се тренира нарочно да дава грешни или непълни твърдения и по този начин да допринася за разпространението на дезинформация, да води до цензура (както например е случаят с DeepSeek и резултати за протестите на площада „Тиананмън“ през 1989 г. [47]) и да му се приписват човешки предразсъдъци (като религия, сексуална ориентация, расизъм, политическа идеология). [48] [49] [50]

Друг голям проблем на генеративния ИИ е свързан с интелектуалната собственост и авторското право. Тук връзката с дезинформацията се отнася за въпросите кой е авторът на дезинформационното съдържание – човекът, накарал машината да създаде невярно съдържание, или машината, която го е създала. Това е важно във връзка с регулациите и евентуална глоба при нарушаване на правилата. Съществува и опасност да се зададе на генеративния модел да направи копие на валидна и вярна информация (било то текст, картина, видео), да се смени, измени, скрие и т.н. малък детайл в нея и дори да се създадат цели сайтове клонинги.

Именно това са част от проблемите, с които ЕС се е заел да разрешава, за да стимулира етичното развитие на ИИ. [51] Първоначално ЕС приема мек подход към ИИ, като през 2019 г. публикува етични принципи за достоверност (Ethics Guidelines for Trustworthy AI) [52] и публични политики и насоки за инвестиции (Policy and investment recommendations). [53] Впоследствие преминава към регулиране на развитието, разпространението и използването на ИИ, и кулминира с приемането на AI Act през 2024 г., част от пакета с мерки на ЕС за хармонизиране на правилата за разработване и използване на ИИ (включващ още AI Innovation Package [54], AI Factories [55] и Coordinated Plan on AI [56]). Комисията на ЕС също така дава старт на доброволната инициатива AI Pact, която има за цел плавно преминаване към спазване на регулаторните мерки на AI Act от всички страни. [57] За ЕС ИИ се дефинира като „машинна система, която е проектирана да функционира с различни нива на автономност и която може да проявява адаптивност след внедряването ѝ, като за ясни и неясни цели, извежда от входа, който получава, как да генерира резултати като прогнози, съдържание, препоръки или решения, които могат да влияят на физически или виртуални среди“. [58]

Регулацията за ИИ, която напълно трябва да влезе в сила през лятото на 2026 г., обаче е много често критикувана най-вече защото нещо толкова

голямо и мощно и с толкова много потенциал (каквото е развитието на ИИ) не може да се регулира чрез технически параметри, на каквито залага документът. [59]

Законът предлага класификация на риска на използваните услуги и задължава разработчиците и разпространителите на ИИ да спазват точни насоки, с които да гарантират прозрачност на услугата, и по този начин да намалят рисковете. [60] Един от основните начини, по които големи доставчици на цифрови услуги (социални мрежи, търсачки и т.н.) могат да работят в посока намаляване на риска (както е упоменато в законодателния документ), е например чрез отбелязване кое съдържание е генерирано от ИИ. По този начин се цели да се увеличи доверието на ползвателите на цифрови услуги, както и да се увеличат възможностите на ползвателите да различават ИИ-генерирана информация и по този начин да се намали рискът от влияние и измама (включително от разпространение на мисинформация). Предстои да се оцени доколко гласуваната регулация ще успее да канализира бързоразвиващото се поле на ИИ, още повече тъй като всяка държава членка ще трябва сама да отговаря за това дали се спазват указанията на закона за ИИ. [61]

В обобщение, когато се говори за ИИ към днешна дата и в по-обширен контекст, по-скоро трябва да се разбират генеративните модели, които използват големи бази данни за машинно обучение и които, чрез общуване с ползвателите на услугата, продължават да се усъвършенстват да решават дадено задание (например генериране на текст, картина, видео). Подобно е и определението за ИИ, с което ЕС борави в опитите си да осигури етично и безопасно развитие на полето.

AI Act и DSA

От Европейската комисия представят нововъведените регулации на онлайн пространството като синхронизирани едни с други. Наблюдават се обаче и несъответствия между AI Act и DSA, които са тревожни именно в контекста на дезинформацията.

DSA полага определени регулаторни рамки, в които доставчиците на онлайн услуги могат да работят. Но генеративните модели не влизат в тези категории, оставяйки встрани от правомощията на DSA. Google например има асистент с ИИ, който резюмира резултата от търсенето, без ползвателят изобщо да отваря самия линк. [62] Това е основната разлика с традиционната търсачка, която е под регулацията на DSA.

В DSA не влизат още комуникационни канали като имейли и приложения за лично общуване (private messaging services), към които може да се присъединят чатботовете, използващи големи езикови модели, когато се използват като канали за общуване в големи онлайн платформи. В такива

случаи например чатботовете могат да са отделна услуга или инструмент, който е част от голяма платформа (обект на DSA регулация), и да се използват за генериране на текст, картина и друг вид директна човешка интервенция. [63]

Затова към този момент AI Act и DSA по-скоро биха работили като паралелни регулации, които имат известни прилики, но могат и да попаднат в конфронтация.

Последно, регулацията за ИИ предвижда най-вече парични глоби, в случай че правилата са нарушени. Докато това би работило при големи платформи и доставчици, по-малките и/или прикрити замесени лица (каквито най-често са разпространителите на дезинформация) трудно биха били обект на глоби. [64] Затова и отговорът не трябва да е само строго регулаторен, а холистичен – обект на колаборации между частния и публичния сектор, гражданския сектор и академичните среди, да включва обучителни програми за критично мислене, етично използване на ИИ и др.

Дискусионни въпроси

Обобщавайки – и дезинформацията, и ИИ са трудно дефинируеми термини, които обаче са широко разпространени, а тяхното взаимодействие се приема за една от големите опасности за демокрациите днес. От една страна, най-големият проблем с дефинирането на дезинформацията се крие в изискването невярната информация да се разпространява с цел да навреди. От друга страна, ИИ е обширно поле, което в частния случай с информационния поток най-често се свързва с машинното обучение и големите езикови модели, на базата на които ИИ може да отговаря на задания, да събира и обобщава информация, но и да съставя изцяло нова информация под формата на текст, картина, видео (генеративни модели). [65]

В този контекст с подобренията и еволюцията на ИИ става все по-трудно да се различава дали дадено съдържание е създадено от човек или машина. Количеството и скоростта на генерираното съдържание са несравними, а цената – изключително ниска в сравнение с традиционните дезинформационни кампании. [66]

Генерирането на текст, картини, видео и т.н. deep fakes никога не е било толкова лесно както е сега, а с всеки ден резултатите се усъвършенстват. Един от основните методи за противодействие на дезинформацията – проверката на факти, няма как да реагира със същата скорост и обем. ИИ може да създава и цели сайтове, които са напълно фалшиви, но ефикасни в ролята си да повлияват мнението на избиратели. [67]

Тук се повдига въпросът и за доверието на гражданите в демократичните процеси – нещо, което дезинформацията цели да ерозира. Отваря се възможност и за нюзрумите да създават цели статии изцяло с помощта

на ИИ [68], което допълнително усложнява качествената журналистика и спомага за вече ниското доверие в медиите.

Някои генеративни модели са научени да не създават нарочно невярна информация (например ChatGPT), но съществуват много модели, които нямат интегрирани етични принципи. Дори и „етични“ модели обаче могат да създават невярна информация поради незнание или халюцинации. Съществува и възможността моделите да се тренират да създават съдържание, което е най-популярното, но не и непременно вярното.

ИИ може да се използва и за създаването на вредни наративи, специфични за дадена държава или тема (например проруски и антиукраински наративи [69]).

Заклучение

Генеративните модели също така са тренирани да говорят различни езици – много по-добре от автоматизирани преводачи като Google Translate, което прави възможно преводът на невярно съдържание още по-лесен и бърз.

Вече съществуват примери как ИИ може да се използва за дезинформация по време на избори [70], да се използва от авторитарни режими и други.

ИИ може да се използва и в борбата с дезинформацията [71]: например за откриването и кодирането на вредно съдържание, както и за разпространението на алтернативни наративи и вярна информация. [72]

Опитите за регулация на цифровия пазар и на ИИ от страна на ЕС са начало, за да се намали възможният ефект от улесняването на разпространението на дезинформация, но както е обобщено по-горе дори в рамките на два законодателни акта съществуват несъответствия и възможни проблеми; само регулация не е достатъчна, а трябва да се поеме по мултидисциплинарния път за цялостно решение.

Цитати и бележки

- [1] BBC News. (2016). *Brexit: The people's verdict*. BBC News. Accessed February 12, 2025. <https://www.bbc.com/news/uk-37995600>.
- [2] Oxford Learner's Dictionaries. (n.d.). *Post-truth*. Oxford University Press. <https://www.oxfordlearnersdictionaries.com/definition/english/post-truth>. Retrieved on 13.02.2025.
- [3] Holan, A. D. (2016, December 13). *2016 lie of the year: 'Fake news'*. Politifact. Accessed February 12, 2025. <https://www.politifact.com/article/2016/dec/13/2016-lie-year-fake-news/>. Retrieved on 13.02.2025.
- [4] Tandoc, E. C., & Seet, S. K. (2022). War of the Words: How Individuals Respond to “Fake News,” “Misinformation,” “Disinformation,” and “Online Falsehoods.” *Jour-*

- nalism Practice*, 18(6), 1503–1519. <https://doi.org/10.1080/17512786.2022.2110929>.
- [5] Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369. <https://doi.org/10.1016/j.jarmac.2017.07.008>.
- [6] McManus, C., & Michaud, C. (2018). *Never mind the buzzwords: Defining fake news and post-truth*. B *Fake News - A Roadmap* (pp. 14–20). Latvia. <https://www.ceeol.com/search/chapter-detail?id=649475>. Retrieved on 02.05.2025.
- [7] Borgesius, F. J. Z., Möller, J., Kruike-meier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & de Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14(1). <https://doi.org/10.18352/ulr.420>.
- [8] Habgood-Coote, J. (2018). Stop talking about fake news! *Inquiry*, 62(9–10), 1033–1065. <https://doi.org/10.1080/0020174X.2018.1508363>.
- [9] Vujnovic, M., & Kruckeberg, D. (2023). Disinformation, misinformation, fake news and their global impact. B Y. R. Kamalipour & J. V. Pavlik (ed.), *Communicating global crises: Media, war, climate, and politics* (pp. 97–115). Rowman & Littlefield.
- [10] OpenAI. (n.d.). *ChatGPT*. OpenAI. <https://openai.com/index/chatgpt/>. Retrieved on 12.02.2025.
- [11] Antoliš, K. (2024). Disinformation supported by artificial intelligence from dynamic research to holistic solutions. *PSPO Journal*. <https://doi.org/10.13165/PSPO-24-35-02>.
- [12] Юркува, М. П. (2022). *Дезинформация онлайн: Стратегии за противодействие в ЕС* (дисертационен труд). Софийски университет „Св. Климент Охридски“, Философски факултет, Катедра „Европеистика“. [Yurukova, M. P. (2022). *Dezinformatsia onlayn: Strategii za protivodeystvie v ES* (disertatsionen trud). Sofiyski universitet „Sv. Kliment Ohridski“, Filosofski fakultet, Katedra „Evropeistika“.]
- [13] *ibid.*
- [14] Юркува, М. П. (2021). Възможности и рискове пред законодателния акт за цифровите услуги по отношение на противодействието на дезинформацията онлайн. *Реторика онлайн*. Retrieved on 13.02.2025. [Yurukova, M. P. (2021). *Vazmozhnosti i riskove pred zakonodatelnia akt za tsifrovite usluzhi po otnoshenie na protivodeystvieto na dezinformatsiyata onlayn*. *Retorika onlayn*.] <https://rhetoric-bg.com/>. Retrieved on 13.02.2025.
- [15] Kaplan, J. (2025). *Meta: More speech, fewer mistakes*. Facebook. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>. Retrieved on 12.02.2025.
- [16] Baker, S. (2025). *Mark Zuckerberg dumps Facebook fact-checker in full view of Elon Musk*. Politico. <https://www.politico.eu/article/mark-zuckerberg-full-elon-musk-dump-facebook-fact-checker/>. Retrieved on 12.02.2025.
- [17] European Commission. (n.d.). *Strengthening online platforms' responsibility: Countering information manipulation*. https://commission.europa.eu/topics/countering-information-manipulation/strengthening-online-platforms-responsibility_bg#ref-%D0%B0%D0%BA%D1%82-%D0%B7%D0%B0-%D0%B8%D0%B7%D0%BA%D1%83%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%B8%D1%8F-

- %D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%B5%D0%BA%D1%82-%D0%B8%D0%B8. Retrieved on 12.02.2025.
- [18] Kroet, C. (2025). *Meta needs to analyse risks if it drops fact-checkers in EU too*, Commission. Euronews. <https://www.euronews.com/next/2025/01/08/meta-needs-to-analyse-risks-if-it-drops-fact-checkers-in-eu-too-commission>. Retrieved on 13.02.2025.
- [19] Booth, R., Milmo, D., & Rankin, J. (2025). *Meta's changes to social media policing will lead to clash with EU and UK, say experts*. The Guardian. Accessed <https://www.theguardian.com/technology/2025/jan/08/metas-changes-to-social-media-policing-will-lead-to-clash-with-eu-and-uk-say-experts>. Retrieved on 12.02.2025.
- [20] Ireton, C., & Posetti, J. (Eds.). (2020). *Journalism, fake news & disinformation: Handbook for journalism education and training*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000265552>. Retrieved on 12.02.2025.
- [21] АЕЖ-България. (2021). *Пътеводител за проверка на факти в дигитална среда*. <https://factcheck.bg/resources/>. Retrieved on 12.02.2025. [AEZh-Bulgaria. (2021). *Patevoditel za proverka na fakti v digitalna sreda*. <https://factcheck.bg/resources/>. Retrieved on 12.02.2025.]
- [22] *ibid.*, стр. 8
- [23] *ibid.*, стр. 8
- [24] *ibid.*, стр. 8
- [25] EEAS. (2023). *Beyond disinformation: What is FIMI?*. Accessed February 12, 2025, from https://www.eeas.europa.eu/eeas/beyond-disinformation-what-fimi_en
- [26] European Commission. (n.d.). *Strengthening online platforms' responsibility: Countering information manipulation*. <https://commission.europa.eu/topics/countering-information-manipulation/strengthening-online-platforms-responsibility>. Retrieved on 12.02.2025.
- [27] *ibid.*
- [28] Юркова, М. П. (2022). *Дезинформация онлайн: Стратегии за противодействие в ЕС* (дисертационен труд). Софийски университет „Св. Климент Охридски“, Философски факултет, Катедра „Европеистика“. [Yurukova, M. P. (2022). *Dezinformatsia onlayn: Strategii za protivodeystvie v ES* (disertatsionen trud). Sofiyski universitet „Sv. Kliment Ohridski“, Filosofski fakultet, Katedra „Evropeistika“.]
- [29] European Commission. (n.d.a). *Digital Services Act*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en. Retrieved on 12.02.2025.
- [30] European Commission. (n.d.b). *Online disinformation*. <https://digital-strategy.ec.europa.eu/en/policies/online-disinformation>. Retrieved on 12.02.2025.
- [31] Traple Konarski Podrecki & Partners. (n.d.). *The European Digital Services Act: A comprehensive overview of the rules*. <https://www.lexology.com/library/detail.aspx?g=ce8a0507-d4bd-42a4-889c-ab16f1a1b209>. Retrieved on 12.02.2025.
- [32] European Commission. (n.d.c). *Digital Markets Act*. European Commission. https://digital-markets-act.ec.europa.eu/index_en. Retrieved on 12.02.2025.

- [33] European Commission. (2018). *Special Eurobarometer 468: Disinformation and online platforms*. <https://europa.eu/eurobarometer/surveys/detail/2183>. Retrieved on 12.02.2025.
- [34] SoftUni. (2025). *AI Basics [Video]*. <https://ai.softuni.bg/trainings/resources/pdf/257/presentation-ai-basics-january-2025>. Retrieved on 12.02.2025.
- [35] Илиева, М. (2024). Изкуствен интелект: правна регулация, отговорност, комуникационни и етични аспекти. *Реторика и комуникации*. бр. 60, 68–86. <https://rhetoric.bg/>. последно посещение на 12.02.2023. [Ilieva, M. (2024). Izkustven intelekt: pravna regulatsia, otgovornost, komunikatsionni i etichni aspekti. Ritorika i komunikatsii. br. 60, 68–86. <https://rhetoric.bg/>. posledno poseshthenie na 12.02.2023.]
- [36] McCarthy, J. (2007). *What is artificial intelligence?* Stanford University. <http://jmc.stanford.edu/articles/whatisai/whatisai.pdf>. Retrieved on 12.02.2025.
- [37] SoftUni. (2025). *AI Basics [Video]*. <https://ai.softuni.bg/trainings/resources/pdf/257/presentation-ai-basics-january-2025>. Retrieved on 12.02.2025.
- [38] Kurzweil, R. (2014). The Singularity is Near. In: Sandler, R.L. (eds) *Ethics and Emerging Technologies*. Palgrave Macmillan, London. https://doi.org/10.1057/9781137349088_26.
- [39] Feuerriegel, S., Hartmann, J., Janiesch, C. et al. (2024). Generative AI. *Business & Information Systems Engineering* 66, 111–126. <https://doi.org/10.1007/s12599-023-00834-7>.
- [40] Kerner, S. M., & Lutkevich, B. (2025). 12 of the best large language models. *TechTarget*. Accessed February 12, 2025, from <https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models>
- [41] Ng, K., Drenon, B., Gerken, T., & Cieslak, M. (2025). *How artificial intelligence could affect the future of work*. *BBC News*. <https://www.bbc.com/news/articles/c5yv5976z9po>. Retrieved on 12.02.2025.
- [42] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article No. 248, 1–38. <https://doi.org/10.1145/3571730>.
- [43] *ibid.*
- [44] Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26), Article eadh1850. <https://doi.org/10.1126/sciadv.adh1850>.
- [45] Feuerriegel, S., Hartmann, J., Janiesch, C. et al. (2024). Generative AI. *Business & Information Systems Engineering* 66, 111–126. <https://doi.org/10.1007/s12599-023-00834-7>. Retrieved on 12.02.2025.
- [46] Chen L, Zaharia M, Zou J (2023) How is chatgpt's behavior changing over time? [arXiv:2307.09009](https://arxiv.org/abs/2307.09009)
- [47] Lu, D. (2025). We tried out DeepSeek: It works well until we asked it about Tiananmen Square and Taiwan. *The Guardian*. <https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan>. Retrieved on 12.02.2025.

- [48] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>
- [49] Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation. <https://doi.org/10.48550/arXiv.2301.01768>. Retrieved on 12.02.2025.
- [50] Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. <https://arxiv.org/abs/2110.01963>
- [51] Madiega, T. (2024). *Artificial intelligence act* (PE 698.792). European Parliamentary Research Service. <https://www.iisf.ie/files/UserFiles/cybersecurity-legislation-ireland/EU-AI-Act.pdf>. Retrieved on 12.02.2025.
- [52] European Commission. (2019a). *Ethics guidelines for trustworthy AI*. Digital Strategy. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>. Retrieved on 12.02.2025.
- [53] European Commission. (2019b). *Policy and investment recommendations for trustworthy artificial intelligence*. Digital Strategy. <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>. Retrieved on 12.02.2025.
- [54] European Commission. (2024). *EU steps up efforts to combat disinformation*. Press corner. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_383. Retrieved on 12.02.2025.
- [55] European Commission. (n.d.d). *AI factories*. Digital strategy. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/ai-factories>. Retrieved on 12.02.2025.
- [56] European Commission. (n.d.e). *Plan for AI*. Digital strategy. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/plan-ai>. Retrieved on 12.02.2025.
- [57] European Commission. (n.d.f). *AI Pact*. Digital strategy. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/ai-pact>. Retrieved on 12.02.2025.
- [58] *ibid*.
- [59] Caroli, L. (2024). *Will the EU AI Act work? Lessons learned from past legislative initiatives & future challenges*. International Association of Privacy Professionals. <https://iapp.org/news/a/will-the-eu-ai-act-work-lessons-learned-from-past-legislative-initiatives-future-challenges>. Retrieved on 12.02.2025.
- [60] Artificial Intelligence Act. (2024). *High-level summary*. Artificial Intelligence Act. Updated on May 30, 2024. <https://artificialintelligenceact.eu/high-level-summary/>. Retrieved on 12.02.2025.
- [61] Fragale, M., & Grilli, V. (2024). Deepfake, deep trouble: The European AI Act and the fight against AI-generated misinformation. *Columbia Journal of European Law*. <https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/>. Retrieved on 12.02.2025.
- [62] Rogers, R. (2024). How to use Google AI and how to turn it off. *Wired*. <https://www.wired.com/story/google-ai-overviews-how-to-use-how-to-turn-off/>
- [63] Calvet-Bademunt, J., & Barata, J. (2024). The Digital Services Act meets the AI Act: Bridging platform and AI governance. *Tech Policy Press*. <https://www.>

- techpolicy.press/the-digital-services-act-meets-the-ai-act-bridging-platform-and-ai-governance/. Retrieved on 12.02.2025.
- [64] Fragale, M., & Grilli, V. (2024). Deepfake, deep trouble: The European AI Act and the fight against AI-generated misinformation. *Columbia Journal of European Law*. <https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/>. Retrieved on 12.02.2025.
- [65] DW Akademie. (2024). Generative AI is the ultimate disinformation amplifier. *DW Akademie*. <https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890>. Retrieved on 12.02.2025.
- [66] Knight, W. (2023). 400 dollars to build an AI disinformation machine. *Wired*. <https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/>. Retrieved on 12.02.2025.
- [67] Cantor, M. (2023). AI-generated news websites: study. *The Guardian*. <https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study>. Retrieved on 12.02.2025.
- [68] Landymore, F. (2023). CNET Is Quietly Publishing Entire Articles Generated By AI. *Futurism*. <https://futurism.com/the-byte/cnet-publishing-articles-by-ai>. Retrieved on 12.02.2025.
- [69] Böswald, L.-M., & Saab, B. (2023). From prompt to problematic. *Disinfo Radar*. Democracy Reporting International. <https://disinforadar.com/wp-content/uploads/2023/04/From-prompt-to-Problematic.pdf>. Retrieved on 12.02.2025.
- [70] Diresta, R., & Miller, C. (2023). How will AI affect the 2024 elections? *Humane Tech Podcast*. <https://www.humanetech.com/podcast/how-will-ai-affect-the-2024-elections-with-renee-diresta-and-carl-miller>. Retrieved on 12.02.2025.
- [71] Li, C., & Callegari, A. (2024). *AI and the fight against online misinformation and disinformation*. World Economic Forum. <https://www.weforum.org/stories/2024/06/ai-combat-online-misinformation-disinformation/>. Retrieved on 12.02.2025.
- [72] Pilati, F., & Venturini, T. (2024). *The use of artificial intelligence in counter-disinformation: A world wide (web) mapping*. SSRN. <https://doi.org/10.2139/ssrn.4855024>.

Библиография

- АЕЖ-България. (2021). *Пътеводител за проверка на факти в дигитална среда*. <https://factcheck.bg/resources/>. Retrieved on 12.02.2025. [АЕЖ-България. (2021). Patevoditel za proverka na fakti v digitalna sreda. <https://factcheck.bg/resources/>. Retrieved on 12.02.2025.]
- Илиева, М. (2024). Изкуствен интелект: правна регулация, отговорност, комуникационни и етични аспекти. *Реторика и комуникации*. бр. 60, 68–86. <https://rhetoric.bg/>. последно посещение на 12.02.2023. [Илиева, М. (2024). Izkustven intelekt: pravna regulatsia, otgovornost, komunikatsionni i etichni aspekti. *Retorika i komunikatsii*. br. 60, 68–86. <https://rhetoric.bg/>. posledno poseshlenie na 12.02.2023.]
- Юркова, М. П. (2021). Възможности и рискове пред законодателния акт за цифровите услуги по отношение на противодействието на дезинформацията

- онлайн. *Реторика онлайн*. <https://rhetoric-bg.com/>. последно посещение на 12.02.2023. [Yurukova, M. P. (2021). Vazmozhnosti i riskove pred zakonodatelnia akt za tsifrovite uslugi po otnoshenie na protivodeystviето na dezinformatsiyata onlayn. Retorika onlayn. <https://rhetoric-bg.com/>. posledno poseshtenie na 12.02.2023.].
- Юркова, М. П. (2022). *Дезинформация онлайн: Стратегии за противодействие в ЕС* (дисертационен труд). Софийски университет „Св. Климент Охридски“, Философски факултет, Катедра „Европеистика“. [Yurukova, M. P. (2022). *Dezinformatsia onlayn: Strategii za protivodeystvie v ES* (disertatsionen trud). Sofiyski universitet „Sv. Kliment Ohridski“, Filosofski fakultet, Katedra „Evropeistika“.]
- Antoliš, K. (2024). Disinformation supported by artificial intelligence from dynamic research to holistic solutions. *PSPO Journal*. <https://doi.org/10.13165/PSPO-24-35-02>.
- Artificial Intelligence Act. (2024). *High-level summary*. Artificial Intelligence Act. Updated on May 30, 2024. <https://artificialintelligenceact.eu/high-level-summary/>. Retrieved on 12.02.2025.
- Baker, S. (2025). *Mark Zuckerberg dumps Facebook fact-checker in full view of Elon Musk*. Politico. <https://www.politico.eu/article/mark-zuckerberg-full-elon-musk-dump-facebook-fact-checker/>. Retrieved on 12.02.2025.
- BBC News. (2016). *Brexit: The people's verdict*. BBC News. <https://www.bbc.com/news/uk-37995600>. Retrieved on 12.02.2025.
- Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. <https://arxiv.org/abs/2110.01963>. Retrieved on 12.02.2025.
- Booth, R., Milmo, D., & Rankin, J. (2025). *Meta's changes to social media policing will lead to clash with EU and UK, say experts*. The Guardian. <https://www.theguardian.com/technology/2025/jan/08/metas-changes-to-social-media-policing-will-lead-to-clash-with-eu-and-uk-say-experts>. Retrieved on 12.02.2025.
- Borgesius, F. J. Z., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & de Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14(1). <https://doi.org/10.18352/ulr.420>.
- Böswald, L.-M., & Saab, B. (2023). From prompt to problematic. *Disinfo Radar*. Democracy Reporting International. <https://disinforadar.com/wp-content/uploads/2023/04/From-prompt-to-Problematic.pdf>. Retrieved on 12.02.2025.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>.
- Calvet-Bademunt, J., & Barata, J. (2024). The Digital Services Act meets the AI Act: Bridging platform and AI governance. *Tech Policy Press*. <https://www.techpolicy.press/the-digital-services-act-meets-the-ai-act-bridging-platform-and-ai-governance/>. Retrieved on 12.02.2025.

- Cantor, M. (2023). AI-generated news websites: study. *The Guardian*. <https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study>. Retrieved on 12.02.2025.
- Caroli, L. (2024). *Will the EU AI Act work? Lessons learned from past legislative initiatives & future challenges*. International Association of Privacy Professionals. <https://iapp.org/news/a/will-the-eu-ai-act-work-lessons-learned-from-past-legislative-initiatives-future-challenges>. Retrieved on 12.02.2025.
- Chen L, Zaharia M, Zou J (2023) How is chatgpt's behavior changing over time? arXiv:2307.09009
- Diresta, R., & Miller, C. (2023). How will AI affect the 2024 elections? *Humane Tech Podcast*. <https://www.humanetech.com/podcast/how-will-ai-affect-the-2024-elections-with-renee-diresta-and-carl-miller>. Retrieved on 12.02.2025.
- DW Akademie. (2024). Generative AI is the ultimate disinformation amplifier. *DW Akademie*. <https://akademie.dw.com/en/generative-ai-is-the-ultimate-disinformation-amplifier/a-68593890>. Retrieved on 12.02.2025.
- EEAS. (2023). *Beyond disinformation: What is FIMI?*. https://www.eeas.europa.eu/eeas/beyond-disinformation-what-fimi_en. Retrieved on 12.02.2025.
- European Commission. (2018). *Special Eurobarometer 468: Disinformation and online platforms*. Retrieved on 12.02.2025.
- European Commission. (2019a). *Ethics guidelines for trustworthy AI*. Digital Strategy. Accessed February 12, 2025, from <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>. Retrieved on 12.02.2025.
- European Commission. (2019b). *Policy and investment recommendations for trustworthy artificial intelligence*. Digital Strategy. Accessed February 12, 2025, from <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>. Retrieved on 12.02.2025.
- European Commission. (2024). *EU steps up efforts to combat disinformation*. Press corner. Accessed February 12, 2025, from https://ec.europa.eu/commission/presscorner/detail/en/ip_24_383. Retrieved on 12.02.2025.
- European Commission. (n.d.). *Strengthening online platforms' responsibility: Countering information manipulation*. https://commission.europa.eu/topics/countering-information-manipulation/strengthening-online-platforms-responsibility_bg#ref-%D0%B0%D0%BA%D1%82-%D0%B7%D0%B0-%D0%B8%D0%B7%D0%BA%D1%83%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%B8%D1%8F-%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%B5%D0%BA%D1%82-%D0%B8%D0%B8. Retrieved on 12.02.2025.
- European Commission. (n.d.a). *Digital Services Act*. Accessed February 12, 2025, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en. Retrieved on 12.02.2025.
- European Commission. (n.d.b). *Online disinformation*. <https://digital-strategy.ec.europa.eu/en/policies/online-disinformation>. Retrieved on 12.02.2025.
- European Commission. (n.d.c). *Digital Markets Act*. European Commission. Retrieved on 12.02.2025.

- European Commission. (n.d.d). *AI factories*. Digital strategy. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/ai-factories>. Retrieved on 12.02.2025.
- European Commission. (n.d.e). *Plan for AI*. Digital strategy. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/plan-ai>. Retrieved on 12.02.2025.
- European Commission. (n.d.f). *AI Pact*. Digital strategy. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/ai-pact>. Retrieved on 12.02.2025.
- Feuerriegel, S., Hartmann, J., Janiesch, C. et al. (2024). Generative AI. *Business & Information Systems Engineering* 66, 111–126. <https://doi.org/10.1007/s12599-023-00834-7>.
- Fragale, M., & Grilli, V. (2024). Deepfake, deep trouble: The European AI Act and the fight against AI-generated misinformation. *Columbia Journal of European Law*. <https://cjel.law.columbia.edu/preliminary-reference/2024/deepfake-deep-trouble-the-european-ai-act-and-the-fight-against-ai-generated-misinformation/>. Retrieved on 12.02.2025.
- Habgood-Coote, J. (2018). Stop talking about fake news! *Inquiry*, 62(9–10), 1033–1065. <https://doi.org/10.1080/0020174X.2018.1508363>.
- Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. <https://doi.org/10.48550/arXiv.2301.01768>.
- Holan, A. D. (2016, December 13). *2016 lie of the year: 'Fake news'*. Politifact. <https://www.politifact.com/article/2016/dec/13/2016-lie-year-fake-news/>. Retrieved on 12.02.2025.
- Ireton, C., & Posetti, J. (Eds.). (2020). *Journalism, fake news & disinformation: Handbook for journalism education and training*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000265552>. Retrieved on 12.02.2025.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article No. 248, 1–38. <https://doi.org/10.1145/3571730>.
- Kaplan, J. (2025). *Meta: More speech, fewer mistakes*. Facebook. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>. Retrieved on 12.02.2025.
- Kerner, S. M., & Lutkevich, B. (2025). 12 of the best large language models. *TechTarget*. Accessed February 12, 2025, from <https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models>
- Knight, W. (2023). 400 dollars to build an AI disinformation machine. *Wired*. Accessed February 12, 2025, from <https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/>
- Kroet, C. (2025). *Meta needs to analyse risks if it drops fact-checkers in EU too, Commission*. Euronews. <https://www.euronews.com/next/2025/01/08/meta-needs-to-analyse-risks-if-it-drops-fact-checkers-in-eu-too-commission>. Retrieved on 12.02.2025.
- Kurzweil, R. (2014). The Singularity is Near. In: Sandler, R.L. (eds) *Ethics and Emerging Technologies*. Palgrave Macmillan, London. https://doi.org/10.1057/9781137349088_26.

- Landymore, F. (2023). CNET Is Quietly Publishing Entire Articles Generated By AI. *Futurism*. Accessed <https://futurism.com/the-byte/cnet-publishing-articles-by-ai>. Retrieved on 12.02.2025.
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369. <https://doi.org/10.1016/j.jarmac.2017.07.008>.
- Li, C., & Callegari, A. (2024). *AI and the fight against online misinformation and disinformation*. World Economic Forum. <https://www.weforum.org/stories/2024/06/ai-combat-online-misinformation-disinformation/>. Retrieved on 12.02.2025.
- Lu, D. (2025). We tried out DeepSeek: It works well until we asked it about Tiananmen Square and Taiwan. *The Guardian*. <https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan>. Retrieved on 12.02.2025.
- Madiega, T. (2024). *Artificial intelligence act* (PE 698.792). European Parliamentary Research Service. Accessed February 12, 2025, from <https://www.iisf.ie/files/UserFiles/cybersecurity-legislation-ireland/EU-AI-Act.pdf>. Retrieved on 12.02.2025.
- McCarthy, J. (2007). *What is artificial intelligence?* Stanford University. <http://jmc.stanford.edu/articles/whatisai/whatisai.pdf>. Retrieved on 12.02.2025.
- McManus, C., & Michaud, C. (2018). *Never mind the buzzwords: Defining fake news and post-truth*. B *Fake News - A Roadmap* (crp. 14–20). Latvia. <https://www.cceol.com/search/chapter-detail?id=649475>
- Ng, K., Drenon, B., Gerken, T., & Cieslak, M. (2025). *How artificial intelligence could affect the future of work*. *BBC News*. <https://www.bbc.com/news/articles/c5yv5976z9po>. Retrieved on 12.02.2025.
- OpenAI. (n.d.). *ChatGPT*. OpenAI. Accessed February 12, 2025, from <https://openai.com/index/chatgpt/>.
- Oxford Learner's Dictionaries. (n.d.). *Post-truth*. Oxford University Press. <https://www.oxfordlearnersdictionaries.com/definition/english/post-truth>. Retrieved on 12.02.2025.
- Pilati, F., & Venturini, T. (2024). *The use of artificial intelligence in counter-disinformation: A world wide (web) mapping*. SSRN. <https://doi.org/10.2139/ssrn.4855024>.
- Rogers, R. (2024). How to use Google AI and how to turn it off. *Wired*. <https://www.wired.com/story/google-ai-overviews-how-to-use-how-to-turn-off/>
- SoftUni. (2025). *AI Basics [Video]*. Accessed <https://ai.softuni.bg/trainings/resources/pdf/257/presentation-ai-basics-january-2025>. Retrieved on 12.02.2025.
- Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26), Article eadh1850. <https://doi.org/10.1126/sciadv.adh1850>.
- Tandoc, E. C., & Seet, S. K. (2022). War of the Words: How Individuals Respond to “Fake News,” “Misinformation,” “Disinformation,” and “Online Falsehoods.”

Journalism Practice, 18(6), 1503–1519. <https://doi.org/10.1080/17512786.2022.2110929>.

Traple Konarski Podrecki & Partners. (n.d.). *The European Digital Services Act: A comprehensive overview of the rules*. <https://www.lexology.com/library/detail.aspx?g=ce8a0507-d4bd-42a4-889c-ab16f1a1b209>. Retrieved on 12.02.2025.

Vujnovic, M., & Kruckeberg, D. (2023). Disinformation, misinformation, fake news and their global impact. В Y. R. Kamalipour & J. V. Pavlik (ed.), *Communicating global crises: Media, war, climate, and politics* (p. 97–115). Rowman & Littlefield.

Тихомира Дончева е докторант в Софийския университет „Св. Климент Охридски“ по направление „Европеистика“ – медийна политика и право на ЕС към катедра „Европеистика“. Завършила е магистратура по „Стратегически комуникации“ в King's College London (2019). Научните ѝ интереси включват медийна политика, достъп до достоверна информация и развитие на медийната грамотност.

Ръкописът е изпратен на 25.04.2025 г.

Рецензиране от двама независими рецензенти: от 01.05.2025 до 02.06.2025 г.

Приемане за публикуване: 03.06.2025 г.

Manuscript was submitted: 25.03.2025.

Double Blind Peer Reviews: from 01.05.2025 till 02.06.2025.

Accepted: 03.06.2025.

Брой 64 на сп. „Реторика и комуникации“ (юли 2025 г.) се издава с финансовата помощ на Фонд научни изследвания, договор № КП-06-НП6/48 от 04 декември 2024 г.

Issue 64 of the Rhetoric and Communications Journal (July 2025) is published with the financial support of the Scientific Research Fund, Contract No. KP-06-NP6/48 of December 04, 2024.