

Наука у доба вештачке интелигенције:

Како вештачка интелигенција мења природу и метод научног истраживања

ОРИГИНАЛ И ВЛАСНИШТВО:

Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research

Issued: October 2024 DES8836_5

ISBN: 978-1-78252-712-1

© The Royal Society

The text of this work is licensed under the terms of the Creative Commons Attribution License which permits unrestricted use, provided the original author and source are credited.

The license is available at:

creativecommons.org/licenses/by/4.0

Images are not covered by this license.

This report and other project outputs can be viewed at:

royalsociety.org/science-in-the-age-of-ai

ПРЕВОДИЛАЦ: СТЕВАН ЈОКИЋ уз помоћ аутоматског преводиоца

Право превода, реализованог Август 2026, је добијено од The Royal Society под условима напред наведене лиценце а биће постављено на сајтове заинтересованих научних друштава у Србији

Напомене преводиоца:

- У тексту нема слика а за сваку слику у оригиналном тексту је дат линк и страница јер немамо право репродукције у преводу на српски језик.
- Сматрам да ће текст бити од користи заинтересованим истраживачима, професорима у школама и свима који желе да сазнају нешто више о вештачкој интелигенцији.
- Свака сугестија у вези текста, посебно научника и стручњака за вештачку интелигенцију, ће бити размотрена и искоришћена за побољшање превода.

САДРЖАЈ

Предговор.....	3
Резиме.....	4
Кључни налази	6
Будућа истраживачка питања	8
Препоруке	9
Увод.....	15
Прво поглавље: Како вештачка интелигенција трансформише научна истраживања	18
Вештачка интелигенција у науци: преглед.....	19
Вештачка интелигенција и методе научног истраживања	22
Вештачка интелигенција и природа научног истраживања	24
Вештачка интелигенција и приступ висококвалитетним подацима	28
Студија случаја: Вештачка интелигенција и дијагноза ретких болести.....	30
Друго поглавље: Интегритет и поузданост истраживања.....	34
Изазови репродуктивности у истраживањима заснованим на вештачкој интелигенцији	35
ВОХ 1:Објашњивост и интерпретабилност.....	37
Баријере које ограничавају репродуктивност	39
Унапређење транспарентности и поузданости	42
ВОХ 2:Робусност и генерализација у машинском учењу.....	42
Треће поглавље: Истраживачке вештине и интердисциплинарност.....	45
Изазови за интердисциплинарност.....	46
Нове истраживачке вештине.....	49
ВОХ 3:Увиди из вежбе црвеног тимирања Краљевског друштва и Хумане интелигенције.....	51
Студија случаја 2: Вештачка интелигенција и наука о материјалима	51
Четврто поглавље: Истраживање, иновације и приватни сектор	55
Променљиви пејзаж технологија вештачке интелигенције у научним истраживањима	57
ВОХ 4:IP Pragmatics 2023: Анализа глобалног патентног пејзажа.....	58
Изазови везани за улогу приватног сектора у науци заснованој на вештачкој интелигенцији	61
ВОХ 5:Улога приватног сектора у патентирању лекова и фармацеутских изума.....	62
Могућности за међусекторску сарадњу	65
Пето поглавље: Истраживачка етика и безбедност вештачке интелигенције.....	66
Разматрање етике вештачке интелигенције у научним истраживањима.....	67
ВОХ 6:Вишејезични језички модели.....	68
Студија случаја 3: Вештачка интелигенција и климатска наука	73
Закључак.....	79
Додаци.....	82-95

ПРЕДГОВОР

Са све већом доступношћу великих база података, нових алгоритамских техника и повећаном рачунарском снагом, вештачка интелигенција (ВИ) постаје устаљени алат који користе истраживачи у различитим научним областима. Сада више него икад, морамо да разумемо обим трансформативног утицаја ВИ на науку и шта научне заједнице треба да ураде да би у потпуности искористиле њене предности.

Овај извештај, „Наука у доба ВИ“, истражује ову тему. Надовезујући се на искуства више од 100 научника који су укључили ВИ у своје радне токове, он се бави тиме како ВИ технологије, као што су дубоко учење или модели великих језика, трансформишу природу и методе научног истраживања. Такође истражује како се појмови интегритета истраживања, истраживачких вештина и истраживачке етике неизбежно мењају – и какве су импликације за будућност науке и научника. Појављују се нове могућности. Студије случаја у овом извештају показују да ВИ побољшава ефикасност, тачност и креативност научника. У више области, примена ВИ отвара нове путеве олакшавајући, на пример, откривање ретких болести или омогућавајући развој одрживијих материјала.

Играјући улогу татора, колеге или асистента, научници користе апликације вештачке интелигенције како би обављали задатке темпом и обимом који су раније били недостижни. Постоји велико узбуђење око синергије између људске интелигенције и вештачке интелигенције и начина на који ово партнерство води ка научној напретку. Међутим, да би се осигурала робусност и ублажиле штете, људска процена и стручност ће и даље бити од највеће важности.

Брза примена вештачке интелигенције у науци такође је представила изазове везане за њену безбедну и ригорозну употребу. Све већи број нерепродуцибилних студија покреће забринутост у вези са робусношћу открића заснованих на вештачкој интелигенцији. Црна кутија и нетранспарентна природа система вештачке интелигенције стварају изазове за верификацију и спољну контролу. Штавише, њено широко распрострањено, али неправедно усвајање покреће етичка питања у вези са њеним утицајем на животну средину и друштво. Па ипак, континуирани напредак у повећању транспарентности и етичкој усклађености система вештачке интелигенције обећава превазилажење ових изазова. У том смислу, извештај позива на уравнотежен приступ који истиче потенцијал вештачке интелигенције у науци, а да притом не изгуби из вида изазове које још увек треба превазићи.

Препоруке нуде пут који користи принципе отворене науке како би се омогућили поуздани научни доприноси вођени вештачком интелигенцијом, истовремено стварајући могућности за дељење ресурса и сарадњу. Такође позивају на политике и праксе које препознају везе између науке и друштва, наглашавајући потребу за етичком вештачком интелигенцијом, једнаким приступом њеним предностима и важност очувања поверења јавности у научна истраживања. Иако је јасно да вештачка интелигенција може значајно помоћи научној напретку, циљ остаје да се осигура да ови продори користе човечанству и планети. Надамо се да ће овај извештај инспирисати актере широм научног екосистема да се укључе у препоруке и раде на будућности у којој можемо да остваримо потенцијал вештачке интелигенције да трансформише науку и користи нашем колективном благостању.

Professor Alison Noble CBE FEng FRS,
Foreign Secretary of the Royal Society and
Chair of the Royal Society Science in the
Age of AI Working Group.

САЖЕТАК

Невиђена брзина и обим напретка вештачке интелигенције (ВИ) последњих година сугеришу да друштво можда проживљава прекретницу. Виралност платформи као што су ChatGPT и Midjourney, које могу да генеришу садржај текста и слика сличан људском, убрзала је интересовање јавности за ову област и подигла забринутост креатора политике у вези начина на који се технологије засноване на ВИ могу интегрисати у шире друштво. Поред тога, коментари истакнутих рачунарских научника и јавних личности у вези са ризицима које ВИ представља за човечанство трансформисали су ову тему у главно политичко питање. За научне истраживаче, ВИ није нова тема и усваја се у неком облику деценијама. Међутим, повећана улагања, интересовање и усвајање у оквиру академских и индустријских истраживања довели су до „револуције дубоког учења“¹ која трансформише пејзаж научних открића.

Омогућене појавом великих база података (на пример, великих и хетерогених облика података прикупљених из телескопа, сателита и других напредних сензора), технике засноване на ВИ помажу у идентификацији нових образаца и односа у великим скуповима података које би иначе било претешко препознати. Ово нуди значајан потенцијал за научна истраживања и подстиче научнике да усвоје сложеније технике које надмашују постојеће методе у њиховим областима. Способност алата вештачке интелигенције да идентификују обрасце из постојећег садржаја и генеришу нова предвиђања такође омогућава научницима да покрећу прецизније симулације и креирају синтетичке податке. Ове симулације, које црпе податке из много различитих извора (потенцијално у реалном времену), могу помоћи доносиоцима одлука да прецизније процене ефикасност потенцијалних интервенција и да се позабаве хитним друштвеним или еколошким изазовима. Могућности вештачке интелигенције за научна истраживања су истакнуте у овом извештају и детаљно истражене кроз три студије случаја о њеној примени за климатску науку, науку о материјалима и дијагнозу ретких болести.

Уз ове могућности, постоје и различити изазови који произилазе из повећаног усвајања вештачке интелигенције. То укључује репродуктивност (у којој други истраживачи не могу да реплицирају експерименте спроведене помоћу алата вештачке интелигенције); интердисциплинарност (где ограничена сарадња између дисциплина које се баве вештачком интелигенцијом и оних које нису из вештачке интелигенције може довести до мање ригорозног усвајања вештачке интелигенције у свим областима); и трошкове заштите животне средине (због велике потрошње енергије потребне за рад велике рачунарске инфраструктуре). Такође постоје све веће препреке за ефикасно усвајање принципа отворене науке због црне кутије система вештачке интелигенције и ограничене транспарентности комерцијалних модела који покрећу истраживања заснована на вештачкој интелигенцији. Штавише, променљиви подстицаји у научном екосистему могу повећати притисак на истраживаче да укључе напредне технике вештачке интелигенције, занемарујући конвенционалније методологије, или да буду „добри у вештачкој интелигенцији“ уместо „добри у науци“².

Ови изазови и потенцијална решења детаљно су описани у овом извештају у поглављима о интегритету истраживања; вештинама и интердисциплинарности; иновацијама и приватном сектору; и истраживачкој етици. Као организација која постоји да би промовисала употребу науке у корист човечанства, ова тема је од великог значаја за Краљевско друштво. Овај извештај, „Наука у доба вештачке интелигенције“, пружа преглед кључних питања којима се треба позабавити како би

¹ Sejnowski T. 2018 *The Deep Learning Revolution*. MIT press

² Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/>. (accessed 7 May 2024)

вештачка интелигенција позитивно трансформисала научни подухват. Његове препоруке, када се узму заједно, требало би да осигурају да примена вештачке интелигенције у научним истраживањима може да достигне свој пуни потенцијал и да помогне у одржавању поверења јавности у науку и интегритета научне методе.

Извештај је вођен радном групом водећих стручњака за вештачку интелигенцију и примењену науку, а информисан је низом активности које је предузело Краљевско друштво. То укључује интервјуе са члановима Краљевског друштва; анализу глобалног патентног пејзажа; преглед историјске литературе; наручену таксономију вештачке интелигенције за научне примене; и неколико радионица на теме које се крећу од модела великих језика до импресивних технологија. Ове активности су у потпуности наведене у додатку. Укупно је више од 100 водећих научних истраживача из различитих дисциплина допринело извештају. Иако извештај покрива нека од критичних подручја везаних за улогу вештачке интелигенције у научним истраживањима, он није свеобухватан и не покрива, на пример, обезбеђивање инфраструктуре за рачунарство високих перформанси, потенцијал опште вештачке интелигенције, нити детаљан преглед нових вештина потребних у различитим индустријама и академским круговима.

Даља истраживачка питања су наведена у наставку. Два програма рада Друштва, на тему „Математичке будућности“³ и „Наука 2040“⁴, детаљније ће истражити релевантне изазове везане за вештине и универзитете.

³ The Royal Society. *Mathematical Futures* Programme. See <https://royalsociety.org/-/media/policy/projects/mathsfutures/mathematical-and-data-education-policy-report.pdf> (accessed 23 April 2024)

⁴ The Royal Society. *Science 2040*. See <https://royalsociety.org/news-resources/projects/science2040/> (accessed 23 April 2024)

Кључни налази

- Поред значајних случајева попут AlphaFold-a, примене вештачке интелигенције могу се наћи у свим STEM областима, са концентрацијом у областима као што су медицина, наука о материјалима, роботика, пољопривреда, генетика и рачунарство. Најистакнутије технике вештачке интелигенције у STEM областима укључују вештачке неуронске мреже, дубоко учење, обраду природног језика и препознавање слика⁵.
 - Висококвалитетни подаци су основа за примене вештачке интелигенције, али истраживачи се суочавају са препрекама везаним за количину, хетерогеност, осетљивост и пристрасност доступних података. Велика количина неких научних података (нпр. прикупљених са телескопа и сателита) може достићи ред петабајта, што отежава постизање циљева као што су дељење података и интероперабилност. Хетерогеност података прикупљених из сензорских података такође представља потешкоће за људску анотацију и стандардизацију, док обука вештачке интелигенције на пристрасним улазима вероватно може довести до пристрасних излаза. С обзиром на ове изазове, кустоси података и менаџери информација су неопходни за одржавање квалитета и решавање ризика повезаних са вештачким генерисањем података, као што су фабриковање података, тровање или контаминација.
 - Индустрија и академске институције унапређују иновације вештачке интелигенције за научна истраживања⁶. У последњој деценији дошло је до пораста броја пријава за патенте везане за вештачку интелигенцију за науку, при чему Кина, Сједињене Америчке Државе, Јапан и Јужна Кореја доминирају по броју поднетих патената. Преглед наручен за овај извештај сугерише да је процена глобалног тржишта вештачке интелигенције (закључно са 2022. годином) приближно 106,99 милијарди фунти⁷.
 - Кина доприноси са приближно 62% патентног пејзажа. Унутар Европе, Велика Британија има други највећи удео патената за вештачку интелигенцију везаних за науку о животу, после Немачке, а академске институције попут Универзитета у Оксфорду, Имperiјал колеџа и Универзитета у Кембриџу истакнуто се налазе међу водећим подносиоцима патената у Великој Британији. Компаније попут Alphabet-a, Siemens-a, IBM-a и Samsung-a изгледа да показују значајан утицај у научним и инжењерским областима.
 - Природа „црне кутије“ и потенцијално власничка природа алата вештачке интелигенције ограничава репродуктивност истраживања заснованих на вештачкој интелигенцији. Препреке попут недовољне документације, ограниченог приступа основним инфраструктурама (нпр. коду, подацима и рачунарској снази) и недостатка разумевања како алати вештачке интелигенције долазе до својих закључака (објашњивости) отежавају независним истраживачима да испитују, верификују и реплицирају експерименте.
- Значајан потенцијал за унапређење открића коришћењем сложених модела дубоког учења такође може подстаћи научнике или финансијере да дају приоритет употреби вештачке интелигенције у односу на ригорозност. Усвајање принципа и пракси отворене науке могло би помоћи у решавању ових изазова и побољшању научног интегритета⁸.

⁵ Berman B, Chubb J, and Williams K, 2024. The use of artificial intelligence in science, technology, engineering, and medicine. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

⁶ Ahmed, N, Wahed, M, & Thompson, N. C. 2023. The growing influence of industry in AI research. *Science*, 379(6635),884-886. (DOI: 10.1126/science.ade2420)

⁷ IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society. See <https://royalsociety.org/newsresources/projects/science-in-the-age-of-ai/>

⁸ UNESCO Recommendation on Open Science. 2021. See: <https://www.unesco.org/en/legal-affairs/recommendationopen-science> (accessed 6 February 2024)

- Интердисциплинарна сарадња је неопходна за премостивање јаза у вештинама и оптимизацију користи вештачке интелигенције у научним истраживањима. Дељењем знања и вештина из међусобних области, сарадња између стручњака за вештачку интелигенцију и предметних области (укључујући истраживаче из уметности, хуманистичких и друштвених наука) може помоћи у стварању ефикаснијих и тачнијих модела вештачке интелигенције. Међутим, ово спречавају изолована истраживачка окружења и структура подстицаја која не награђује интердисциплинарну сарадњу у смислу доприноса напредовању у каријери.
- Генеративни алати вештачке интелигенције могу помоћи у унапређењу научних истраживања. Они обећавају убрзавање рутинских научних задатака, као што су обрада неструктурираних података, решавање сложених изазова кодирања или подршка вишејезичном превођењу академских чланака. Поред тога, може постојати место за моделе генерисања текста који ће се користити за академске и неакадемске писане задатке, са потенцијалним импликацијама на научну комуникацију и процену истраживања. Као одговор на то, финансијери и академске институције постављају норме како би спречили непожељну употребу^{9,10}

⁹ Kaiser J. 2023. Funding agencies say no to AI peer review. *Science*. 14 July 2023 .See: <https://www.science.org/content/article/science-funding-agencies-say-no-using-ai-peer-review> (accessed 23 January 2024)

¹⁰ Harker J. 2023. Science Journals set new authorship guidelines for AI-generated text. National Institute of Environmental Health Sciences. See <https://factor.niehs.nih.gov/2023/3/feature/2-artificial-intelligence-ethics>.(accessed 23 January 2024)

Питања за будућа истраживања

Следеће теме су се појавиле у истраживачким активностима као кључна разматрања за будућност вештачке интелигенције у науци:

1. **Вештачка интелигенција и рачунарске инфраструктуре за науку:** Како се радна оптерећења вештачке интелигенције могу оптимизовати да би се искористио пун потенцијал хетерогених рачунарских инфраструктура у научном истраживању, узимајући у обзир различите захтеве различитих научних области?
2. **Вештачка интелигенција и малобројни подаци:** Које су импликације све веће употребе вештачке интелигенције за истраживаче којима су доступни само малобројни подаци? Како се технике вештачке интелигенције могу ефикасно користити за проширивање малих скупова података у сврху обуке? Који компромиси постоје између смањења величине модела и очувања перформанси када се примене на сценарије са малобројним подацима?
3. **Вештачка интелигенција и неједнакости у научном систему:** Које препреке постоје у обезбеђивању једнаког приступа технологијама вештачке интелигенције у недовољно заступљеним заједницама? Како се вештачка интелигенција може користити за проширење учешћа међу научним и стручним заједницама, укључујући недовољно заступљен број научника и ненаучну јавност?
4. **Вештачка интелигенција и интелектуална својина:** Који су улази система вештачке интелигенције (скупови података, алгоритми или излази) кључни за заштиту интелектуалне својине и на које начине то утиче на примену принципа отворене науке у науци?
5. **Вештачка интелигенција и будућност вештина за науку:** Како се захтеви за вештинама у научним истраживањима мењају са све већом интеграцијом вештачке интелигенције? Које компетенције ће бити неопходне за истраживаче у будућности и који напори су потребни за промоцију писмености у области вештачке интелигенције у различитим научним дисциплинама?
6. **Вештачка интелигенција и будућност научне комуникације:** Како се пејзаж академске и научне комуникације развија са интеграцијом технологија вештачке интелигенције? Како се вештачка интелигенција може искористити за побољшање преношења знања, вишејезичности и мултимодалности у научним резултатима?
7. **Вештачка интелигенција и еколошка одрживост:** Какву улогу вештачка интелигенција може играти у промоцији одрживих пракси унутар научне заједнице? Како се алгоритми вештачке интелигенције могу оптимизовати како би се побољшала енергетска ефикасност моделирања животне средине и допринело одрживим праксама у областима као што су климатска наука, екологија и праћење животне средине?
8. **Стандарди вештачке интелигенције и научна истраживања:** Како стандарди вештачке интелигенције могу помоћи у решавању изазова репродуктивности или интероперабилности у научним истраживањима заснованим на вештачкој интелигенцији? Како научна заједница може допринети успостављању стандарда вештачке интелигенције?

ПРЕПОРУКЕ

ОБЛАСТ ЗА АКЦИЈУ: УНАПРЕДИТИ ПРИСТУП ОСНОВНИМ ИНФРАСТРУКТУРАМА И АЛАТИМА

ПРЕПОРУКА 1.

Владе, финансијери истраживања и програмери вештачке интелигенције требало би да побољшају приступ основним инфраструктурама вештачке интелигенције

Приступ рачунарским ресурсима био је кључан за велике научне продоре, као што је савијање протеина помоћу програма AlphaFold. Упркос томе, рачунарска снага и инфраструктура података за истраживање вештачке интелигенције нису подједнако доступне нити распоређене међу истраживачким заједницама¹¹. Научницима из различитих дисциплина потребан је приступ инфраструктури како би усвојили сложеније технике вештачке интелигенције, обрађивали веће количине и типове података и осигурали квалитет у истраживањима заснованим на вештачкој интелигенцији.

Предлози за побољшање приступа укључивали су институције које спонзоришу приступ суперрачунарству¹² и успостављање регионалних центара, – сличних ЦЕРН-у, за вештачку интелигенцију¹³. Шири приступ може проширити предности вештачке интелигенције на већи број дисциплина, побољшати конкурентност истраживача ван индустрије и допринети ригорознијој науци омогућавањем репродуктивности у великим размерама. Проширење приступа рачунарству такође мора бити информисано најбољим праксама еколошки одрживе рачунарске науке (ESCS), укључујући мерење и извештавање о утицајима на животну средину¹⁴.

Акције за побољшање приступа инфраструктурама и алатима вештачке интелигенције могу укључивати:

1. Финансијери, индустријски партнери и истраживачке институције са рачунарским објектима које активно деле основне инфраструктуре вештачке интелигенције као што су високоперформансна рачунарска снага и ресурси података.
2. Релевантне заинтересоване стране (нпр. владине агенције, истраживачке институције, индустрија и међународне организације) које обезбеђују приступ висококвалитетним скуповима података и интероперабилним инфраструктурама података у свим секторима и регионима. Ово би

¹¹ Technopolis Group, Alan Turing Institute. 2022. Review of Digital Research Infrastructure Requirements for AI. See: https://www.turing.ac.uk/sites/default/files/2022-09/ukri-requirements-report_final_edits.pdf (accessed February 6 2024)

¹² UKRI. Transforming our world with AI. See: <https://www.ukri.org/publications/transforming-our-world-with-ai/> (accessed 6 February 2024)

¹³ United Nations. 2023 Interim Report: Governing AI for Humanity. See: https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf (accessed 6 February 2024)

¹⁴ Lannelongue, L, *et al.* 2023. Greener principles for environmentally sustainable computational science. Nat Comput Sci3, 514–521. (<https://doi.org/10.1038/s43588-023-00461-y>)

могло да укључи унапређење приступа осетљивим подацима путем технологија за побољшање приватности и поузданих истраживачких окружења¹⁵.

3. Финансијери истраживања који подржавају стратегије за праћење и ублажавање утицаја на животну средину повезаног са повећаним рачунарским захтевима и унапређење принципа пропорционалности енергије у апликацијама вештачке интелигенције¹⁶.

PREPORUKA 2

Финансијери и програмери вештачке интелигенције требало би да дају приоритет приступачности и употребљивости алата вештачке интелигенције развијених за научна истраживања

Приступ вештачкој интелигенцији не гарантује њену смислену и одговорну употребу. Сложени и високо ефикасни алати и методе вештачке интелигенције могу бити изазовни за истраживаче који нису из области вештачке интелигенције да их усвоје и ефикасно користе¹⁷. Слично томе, потребне су нове вештине током целог животног циклуса вештачке интелигенције, као што су научници за податке који разумеју важност метаподатака и курирања података или инжењери који су упознати са програмирањем помоћу графичких процесора за обраду слика.

Предузимање корака за побољшање употребљивости алата заснованих на вештачкој интелигенцији (нпр. софтверске апликације, библиотеке, API-ји или општи системи вештачке интелигенције) стога треба да укључује комбинацију механизма који чине вештачку интелигенцију разумљивом за стручњаке који нису из области вештачке интелигенције и граде њихов капацитет за одговорно коришћење вештачке интелигенције. На пример, обука треба да осигура да сваки научник буде у стању да препозна када му је потребна специјализована стручност у области података или програмирања у својим тимовима, или када би употреба сложених и непрозирних техника вештачке интелигенције могла да угрози интегритет и квалитет резултата.

Побољшање употребљивости такође може побољшати улогу научника који нису из области вештачке интелигенције као ко-дизајнера¹⁸ – за разлику од пасивних корисника – који могу да осигурају да алати вештачке интелигенције задовољавају потребе научне заједнице. Стварање услова за дизајн кода захтева премошћавање дисциплинарног јаза између стручњака за вештачку интелигенцију и доменских стручњака кроз развој заједничких језика, начина рада и алата.

Акције за побољшање употребљивости алата вештачке интелигенције могу да укључују:

1. Истраживачке институције и центри за обуку успостављају наставне планове и програме за писменост у области вештачке интелигенције у свим научним областима како би изградиле капацитет истраживача да разумеју могућности, ограничења и адекватност алата заснованих на вештачкој интелигенцији у оквиру својих области и истраживачких контекста.
2. Истраживачке институције и центри за обуку успостављају свеобухватне наставне планове и програме за писменост у вези са подацима прилагођене специфичним потребама примена

¹⁵ The Royal Society. 2023 Privacy Enhancing Technologies. See <https://royalsociety.org/topics-policy/projects/privacyenhancing-technologies/> (accessed 21 December 2023).

¹⁶ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023).

¹⁷ Cartwright H. 2023 Interpretability: Should – and can – we understand the reasoning of machine-learning systems? In: OECD (ed.) *Artificial Intelligence in Science*. OECD. (<https://doi.org/10.1787/a8d820bd-en>)

¹⁸ UKRI. Trustworthy Autonomous Systems Hub. Developing machine learning models with codesign: how everyone can shape the future of AI. See: <https://tas.ac.uk/developing-machine-learning-models-with-codesign-how-everyone-can-shape-the-future-of-ai/> (accessed 7 March 2023)

вештачке интелигенције у научним истраживањима. Ово подразумева изградњу капацитета за управљање, прикупљање и управљање подацима, као и примену принципа података као што су FAIR (Findable, Accessible, Interoperable, and Reusable - Пронађиво, Приступачно, Интероперабилно и Поновно употребљиво) и CARE (Collective benefit, Authority to control, Responsibility, and Ethics - Колективна корист, Овлашћење за контролу, Одговорност и Етика)¹⁹.

3. Финансијери истраживања и програмери вештачке интелигенције улажу у стратегије које побољшавају разумевање и употребљивост вештачке интелигенције за нестручњаке за вештачку интелигенцију, са фокусом на сложене и непрозирне моделе²⁰. Ово може да укључује даља истраживања о објашњивим вештачким интелигенцијама (XAI) специфичним за домен, или приступачним алатима вештачке интелигенције који побољшавају приступ у истраживачким окружењима са ограниченим ресурсима²¹.

4. Истраживачке институције, финансијери истраживања и научни часописи који имплементирају механизме за олакшавање преношења знања из домена и смислене сарадње између дисциплина. Ово захтева комбинацију интердисциплинарне обуке, менторства, публикација и финансирања (нпр. путем тела као што је Споразум о међусаветском овлашћењу UKRI-ја који регулише интердисциплинарне предлоге истраживања)²²

ПРЕПОРУКА 3

ОБЛАСТ ЗА ДЕЛОВАЊЕ: ИЗГРАДИТИ ПОВЕРЕЊЕ У ИНТЕГРИТЕТ И КВАЛИТЕТ НАУЧНИХ РЕЗУЛТАТА ЗАСНОВАНИХ НА ВЕШТАЧКОЈ ИНТЕЛЕКТУРИ

Све већи број нерепродуцибилних студија заснованих на вештачкој интелигенцији и машинском учењу (МУ) изазива забринутост у вези са поузданошћу открића заснованих на вештачкој интелигенцији^{23,24}. Међутим, научници се суочавају са изазовима у побољшању репродуктивности свог рада заснованог на вештачкој интелигенцији. То укључује недовољну документацију објављену о методама, коду, подацима или рачунарским окружењима²⁵; ограничен приступ рачунарству за валидацију сложених модела МУ²⁶; и ограничене награде за имплементацију пракси отворене науке²⁷. Ово представља ризике не само за науку, већ и за друштво, ако примена непоузданих или неисплативих резултата заснованих на вештачкој

¹⁹ Global Indigenous Data Alliance. Care Principles for Indigenous Data Governance. See <https://www.gida-global.org/care> (accessed 21 December 2023)

²⁰ Szymanski M, Verbert K, Vanden Abeele V. 2022. Designing and evaluating explainable AI for non-AI experts: challenges and opportunities. In Proceedings of the 16th ACM Conference on Recommender Systems (<https://doi.org/10.1145/3523227.3547427>)

²¹ Korot E *et al.* 2021 Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell.* 3, 288–298. (<https://doi.org/10.1038/s42256-021-00305-2>)

²² UKRI. Get Support For Your Project: If your research spans different disciplines. See: <https://www.ukri.org/apply-forfunding/how-to-apply/preparing-to-make-a-funding-application/if-your-research-spans-different-disciplines/> (accessed 13 December 2023)

²³ Haibe-Kains B *et al.* 2020 Transparency and reproducibility in artificial intelligence. *Nature.* 586, E14–E16. (<https://doi.org/10.1038/s41586-020-2766-y>)

²⁴ Kapoor S and Narayanan A. 2023 Leakage and the reproducibility crisis in machine-learning-based science. *Patterns.* 4(9) (<https://doi.org/10.1016/j.patter.2023.100804>)

²⁵ Pineau, J, *et al.* 2021. Improving reproducibility in machine learning research (a report from the Neurips 2019 Reproducibility program).” *Journal of Machine Learning Research* 22.164.

²⁶ Bommasani *et al.* 2021. On the opportunities and risks of foundation models. See: <https://crfm.stanford.edu/assets/report.pdf> (accessed 21 March 2024)

²⁷ UK Parliament, Reproducibility and Research Integrity – Report Summary See: <https://publications.parliament.uk/pa/cm5803/cmselect/cmsctech/101/summary.html> (accessed 7 February 2024)

интелигенцији доведе до штетних исхода²⁸ Да би се решили ови изазови, вештачка интелигенција у науци може имати користи од праћења принципа и пракси отворене науке. На пример, УНЕСКО-ва препорука о отвореној науци²⁹ нуди релевантне смернице за побољшање научне ригорозности, уз напомену да не постоји универзални приступ за практиковање отворености у свим секторима и регионима. Ово се добро слаже са растућом тенденцијом усвајања „постепених“ отворених модела који упарују отворено објављивање модела и података са имплементацијом детаљних смерница и заштитних ограда за веродостојне ризике³⁰.

Принципи отворене науке такође могу допринети праведнијем приступу предностима вештачке интелигенције и изградњи капацитета ширег круга стручњака да допринесу њеној примени у науци. Овим принципом се пружа могућност укључивања: научника који су ван ове области, власника података који могу бити од користи у разним истраживањима и заинтересоване ненаучне јавности. У сваком случају потребан је даљи интензиван рад како би се разумеле интеракције између отворене науке и вештачке интелигенције које се сматрају корисним за науку, као и како да се минимизирају безбедносни ризици који произилазе из објављивања модела отворене науке и података.

Акције за промоцију усвајања отворене науке у науци заснованој на вештачкој интелигенцији могу укључивати:

1. Финансијере истраживања и истраживачке институције које подстичу усвајање принципа и пракси отворене науке ради побољшања репродуктивности истраживања заснованих на вештачкој интелигенцији. На пример, доделом средстава за отворену науку и обуку за вештачку интелигенцију, захтевањем коришћења контролних листа репродуктивности³¹ и протокола за дељење података као дела пријава за грантове, или подржавањем развоја стандарда репродуктивности специфичних за заједницу и одговарајућу област (нпр. TRIPOD-AI³²).
2. Истраживачке институције и часописи награђују и признају праксе отворене науке у могућностима за напредовање у каријери. На пример, промовисањем дисеминације неуспелих резултата, прихватањем извештаја о претходној регистрацији и регистрованим извештајима као излазних резултата или признавањем објављивања скупова података и документације као релевантних публикација за напредовање у каријери.
3. Финансијери истраживања, истраживачке институције и актери из индустрије подстичу међународну сарадњу улагањем у инфраструктуру, алате и праксе отворене науке. На пример, улагањем у отворене репозиторијуме који омогућавају дељење скупова података, верзија софтвера и токова рада, или подржавањем развоја документације која је контекстуално одговарајућа и омогућава локалну адаптацију модела вештачке интелигенције у различитим истраживачким окружењима. Ово последње може допринети и укључивању недовољно заступљених

²⁸Sambasivan, N, *et al* 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

²⁹ UNESCO Recommendation on Open Science. 2021. See: <https://www.unesco.org/en/legal-affairs/recommendationopen-science> (accessed 6 February 2024)

³⁰ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

³¹ McGill School of Computer Science. The Machine Learning Reproducibility Checklist v2.0. See: <https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf> (accessed 21 December 2023).

³² Collins G *et al*. 2021 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*, 11(7), e048008. (<https://doi.org/10.1136/bmjopen-2020-048008>)

истраживачких заједница и научника који раде у одговарајућим контекстима али са ниским ресурсима.

4. Релевантни креатори политике разматрају начине одвраћања од развоја затворених екосистема за вештачку интелигенцију у науци, на пример, прописивањем одговорног објављивања референтних вредности, података за обуку и методологија које се користе у истраживањима која води индустрија.

ПРЕПОРУКА 4

Научне заједнице треба да изграде капацитете за надгледање система вештачке интелигенције који се користе у науци и да обезбеде њихову етичку употребу за јавно добро.

Примена вештачке интелигенције у научним областима захтева пажљиво разматрање потенцијалних ризика и случајева злоупотребе. То може укључивати утицај пристрасности података³³, тровања података³⁴, ширења научних дезинформација^{35,36} и злонамерне промене модела вештачке интелигенције³⁷. Поред тога, интензивна природа вештачке интелигенције (нпр. у смислу енергије, података и људског рада) покреће етичка питања у вези са мером у којој вештачка интелигенција коју користе научници може ненамерно допринети штети по животну средину и друштво.

Етичке бриге су погоршане неизвесношћу која окружује ризике вештачке интелигенције. Од краја 2023. године, јавне дебате о безбедности вештачке интелигенције нису коначно дефинисале улогу научника у праћењу и ублажавању ризика у њиховим областима. Штавише, различити нивои техничке стручности у вези са вештачком интелигенцијом међу стручњацима у области и недостатак стандардизованих метода за спровођење процена утицаја на етику ограничавају способност научника да обезбеде ефикасан надзор³⁸. Други фактори укључују ограничену транспарентност комерцијалних модела, непрозирну природу система машинског учења и како злоупотреба пракси отворене науке може повећати безбедносне ризике^{39,40}.

³³ Arora, A, Barrett, M, Lee, E, Oborn, E and Prince, K 2023 Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, **33**. (<https://doi.org/10.1016/j.infoandorg.2023.100478>)

³⁴ Verde, L., Marulli, F. and Marrone, S., 2021. Exploring the impact of data poisoning attacks on machine learning model reliability. *Procedia Computer Science*, **192**. 2624-2632. (<https://doi.org/10.1016/j.procs.2021.09.032>)

³⁵ Truhn D, Reis-Filho J.S. & Kather J.N. 2023 Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat Med* **29**, 2983–2984. (<https://doi.org/10.1038/s41591-023-02594-z>)

³⁶ The Royal Society. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation. See <https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/>

³⁷ Kazim, E and Koshiyama, A.S 2021 A high-level overview of AI ethics. *Patterns*, **2**. (<https://doi.org/10.1016/j.patter.2021.100314>)

³⁸ Wang H *et al.* 2023 Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (<https://doi.org/10.1038/s41586-023-06221-2>)

³⁹ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

⁴⁰ Vincent J. 2023 OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. *The Verge*. See <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskeverinterview> (accessed 21 December 2023).

Како се вештачка интелигенција даље интегрише у науку, потребни су механизми за њено осигурање⁴¹ како би се одржало поверење јавности у њу и осигурао одговоран научни напредак који користи човечанству. Сарадња између стручњака за вештачку интелигенцију, стручњака из области и истраживача из хуманистичких и научних, технолошких, инжењерских, уметничких и математичких (STEAM) дисциплина може побољшати способност научника да надгледају системе вештачке интелигенције и предвиде штету⁴². Слично томе, ангажовање заједница представљених у скуповима података за обуку вештачке интелигенције или одсутних из њих може побољшати тренутно разумевање могућих ризика и штетности истраживачких пројеката заснованих на вештачкој интелигенцији.

Акције за подршку етичкој примени вештачке интелигенције у науци могу укључивати:

1. Финансијери истраживања и институције које улажу у рад који операционализује и успоставља таксономије специфичне за домен⁴³ ризика вештачке интелигенције у науци, посебно осетљивим областима (нпр. хемијска и биолошка истраживања).
2. Финансијери истраживања, истраживачке институције, актери из индустрије и релевантне научне заједнице које прихватају широко доступне етичке оквире за вештачку интелигенцију, као што је дато у Препоруци УНЕСКО-а о етици вештачке интелигенције⁴⁴ или Етичким смерницама ОЕЦД-а за вештачку интелигенцију⁴⁵, и примењују праксе које спајају отворену науку са заштитним мерама против потенцијалних ризика.
3. Финансијери, истраживачке институције и центри за обуку који пружају обуку о етици вештачке интелигенције и граде капацитете научника за спровођење активности предвиђања (нпр. Хоризонтално скенирање вештачке интелигенције је употреба вештачке интелигенције за систематско праћење, филтрирање и анализу екстерних података.), тестирање пре распоређивања (AI *red teaming* деловање је процес супарничког тестирања где стручњаци симулирају нападе из стварног света како би открили безбедносне пропусте, кршења безбедности и ризике од злоупотребе у AI системима пре него што се они примене.) или процене етичког утицаја модела вештачке интелигенције како би се идентификовали релевантни ризици и заштитне ограде повезане са њиховом облашћу.
4. Финансијери истраживања, истраживачке институције и центри за обуку који подржавају развој интердисциплинарних и партиципативних приступа ревизији безбедности, обезбеђујући укључивање научника који се баве вештачком интелигенцијом и не, као и погођених заједница у евалуацију примене вештачке интелигенције за научна истраживања.

⁴¹ Brennan, J. 2023. AI assurance? Assessing and mitigating risks across the AI lifecycle. Ada Lovelace Institute. See <https://www.adalovelaceinstitute.org/report/risks-ai-systems/> (accessed 30 September 2023)

⁴² The Royal Society. 2023. Science in the metaverse: policy implications of immersive technology. See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/>

⁴³ Weidinger L, et al. 2022 Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 214-229. (<https://doi.org/10.1145/3531146.3533088>)

⁴⁴ UNESCO. 2022. Recommendation on the ethics of artificial intelligence. See: <https://www.unesco.org/en/artificialintelligence/recommendation-ethics> (accessed 5 March 2024)

⁴⁵ OECD. Ethical guidelines for artificial intelligence. See: <https://oecd.ai/en/catalogue/tools/ethical-guidelines-forartificial-intelligence> (accessed 5 March 2024)

УВОД

Обим извештаја

Документ, *Наука у доба вештачке интелигенције*, истражује како вештачка интелигенција трансформише природу и методе научног истраживања. Фокусира се на утицај метода дубоког учења, примена генеративне вештачке интелигенције, и истражује међусекторска разматрања око интегритета самог истраживања, вештина и етике. Док вештачка интелигенција управо трансформише широк спектар области – укључујући друштвене и хуманистичке науке – дотле овај извештај посебну пажњу посвећује конкретним примерима из физичке и биолошке науке.

Разматрана су, између осталих, и следећа питања:

- Како технологије засноване на вештачкој интелигенцији трансформишу методе и природу научних истраживања?
- Које су могућности, ограничења и ризици ових технологија за научна истраживања?
- Како релевантне заинтересоване стране (владе, универзитети, индустрија, финансијери истраживања, итд.) могу најбоље подржати развој, усвајање и употребу технологија заснованих на вештачкој интелигенцији у научним истраживањима?

Свако поглавље пружа доказе прикупљене из интервјуа, округлих столова, радионица и наручених истраживања спроведених за овај извештај како би се одговорило на ова питања.

Резултати су представљени на следећи начин:

Поглавље 1 пружа дескриптивни преглед како недавни развој вештачке интелигенције (посебно у машинском учењу (МУ), дубоким неуронским мрежама и обради природног језика) мења методе, процесе и праксе у научним истраживањима.

• **Поглавље 2** детаљно описује кључне изазове за интегритет истраживања заснованим на вештачкој интелигенцији. Бави се питањима транспарентности модела и скупова података вештачке интелигенције, објашњивости и интерпретабилности, као и препрекама за проверу репродуктивности резултата.

• **Поглавље 3** бави се интердисциплинарном сарадњом и новим истраживачким вештинама у научним истраживањима вођеним вештачком интелигенцијом. Испитује изазове као што су изоловане академске културе и инфраструктуре података, истражује могућности за сарадњу између различитих области и карактерише потребу за програмима усавршавања и обуке у области великих база података, вештачке интелигенције и етике.

• **Поглавље 4** бави се растућом улогом приватног сектора у истраживањима заснованим на вештачкој интелигенцији. Разматра прилике и изазове везане за утицај приватног сектора на јавни сектор, као и примере међусекторске сарадње.

• **Поглавље 5** се бави истраживачком етиком и безбедношћу у истраживањима заснованим на вештачкој интелигенцији. Истиче потребу за надзором како би се спречиле штетне последице везане за етичке изазове, безбедност и безбедносне ризике. Примери укључују пристрасност података, халуцинације или пренамену скупова података и модела са злонамерном намером.

Извештај такође укључује студије случаја о примени вештачке интелигенције за климатску науку, науку о материјалима и дијагнозу ретких болести. Ови случајеви такође истражују изазове везане за приступ истраживача различитим подацима (Студија случаја 1), импликације аутоматизације у лабораторији (Студија случаја 2) и нова разматрања истраживачке етике у применама вештачке интелигенције у науци (Студија случаја 3).

Циљне публике овог извештаја

Овај извештај би мога бити користан:

- Научницима и финансијерима истраживања који се на неки начин сналазе у променљивој улози вештачке интелигенције у науци. Извештај им нуди преглед могућности и изазова повезаних са интеграцијом све сложенијих техника у научна истраживања.
- Стручњацима и програмерима за вештачку интелигенцију извештај омогућује интердисциплинарну сарадњу са стручњацима из научних области и јачање међусекторске сарадње.
- Доносиоцима политика и регулаторима, укљученим у обликовање стратегија за вештачку интелигенцију и податке, докази прикупљени у овом извештају могу бити искоришћени при развоју стратегија које промовишу одговоран развој вештачке интелигенције, обрађују етичка питања, и подржавају научни напредак.
- Широј јавности, која жели да разуме будуће правце и примене вештачке интелигенције, извештај даје допринос информисању о могућностима и изазовима повезаним са усвајањем вештачке интелигенције, као и ширим друштвеним импликацијама унапређења ове технологије. Читаоци не морају имати техничко знање о вештачкој интелигенцији да би с разумевањем могли да прочитају овај извештај.

Речник кључних појмова

Извештај се ослања на концепте из области науке о подацима и вештачке интелигенције⁴⁶. Дат је преглед кључних појмова коришћених у извештају.

Вештачка интелигенција (ВИ- Artificial intelligence (AI)):

Развој и проучавање машина способних да обављају задатке који су конвенционално захтевали људске когнитивне способности. Обухвата различите аспекте интелигенције, као што су расуђивање, доношење одлука, учење, комуникација, решавање проблема и физичко кретање. ВИ налази широку примену у свакодневним технологијама попут виртуелних асистената, претраживача, навигационих система и онлајн банкарства.

Вештачке неуронске мреже (ВНМ- Artificial neural networks (ANNs)): Системи вештачке интелигенције, инспирисани структуром биолошког мозга, се састоје од међусобно повезаних рачунарских јединица (неурона) способних за пренос података између слојева неурона. Препознајемо их при реализацији задатака попут препознавања лица и гласа, са више слојева који доприносе способностима решавања постављених проблема. Видите такође „дубоко учење“.

Дубоко учење (ДУ- Deep learning (DL)): Облик машинског учења који користи рачунарске структуре познате као „вештачке неуронске мреже“ за аутоматско препознавање образаца у подацима и производњу релевантних резултата. Инспирисани биолошким мозгом, модели дубоког учења су успешни у сложеним задацима као што су препознавање слика и говора, а налазе и апликације у улози гласовних асистената и аутономних возила. Видите „Вештачке неуронске мреже“.

Основни модел (Foundation model): Модел машинског учења обучен на опсежним подацима, прилагодљив за различите примене. Уобичајени примери укључују велике језичке моделе, који служе као основа за разне вештачке интелигенције, попут четботова.

Генеративна вештачка интелигенција (Generative AI): Системи вештачке интелигенције који генеришу нови текст, слике, аудио или видео као одговор на кориснички унос користећи технике машинског учења. Ови системи, често користећи генеративне супарничке мреже (ГАН- General adversarial networks (GANs)), креирају излазе који веома подсећају на медије које су створили

⁴⁶ The Alan Turing Institute. Defining data science and AI. See: <https://www.turing.ac.uk/news/data-science-and-aiglossary> (accessed 1 March 2024)

људи, што резултира излазима који се често не разликују од медија које су створили људи. Видите „Генеративне супарничке мреже“.

Опште супарничке мреже (ГАН- General adversarial networks (GANs)): Техника машинског учења која производи реалистичне синтетичке податке, попут дубоко лажних слика, неразлучиве од својих података за обуку. Састоје се од генератора, који креира лажне податке, и дискриминатор који их процењује у односу на стварне податке, помажући тако генератору да се побољша све док дискриминатор не дође у ситуацију да није у могућности да разликује стварне од лажних података.

Човек у петљи (HITL- Human-in-the-loop (HITL)): Хибридни систем који се састоји од људске и вештачке интелигенције који омогућава људску интервенцију, као што је обука или фино подешавање алгоритама, како би се побољшао излаз система. Комбиновање снаге људског расуђивања и машинских могућности може надокнадити ограничења оба.

Велики језички модели (LLM- Large language models (LLM)): Основни модели обучени на опсежним текстуалним подацима за обављање задатака везаних за језик, укључујући четботове и генерисање текста. Они су део ширег поља истраживања које се назива обрада природног језика и обично су много једноставнијег дизајна од мањих, традиционалних језичких модела.

Машинско учење (МУ- Machine learning (ML)): Област вештачке интелигенције која укључује алгоритме који уче обрасце из података и примењују те налазе како би правили предвиђања или нудили корисне резултате. Омогућава решавање задатака попут превођења језика, медицинске дијагнозе и роботске навигације анализирајући узорке података ради побољшања перформанси током времена.

Технологије за побољшање приватности (ТПП- Privacy-enhancing technologies (PETs)): Кровни термин који покрива широк спектар технологија и приступа који могу помоћи у ублажавању ризика по безбедност података и приватност⁴⁷.

Синтетички подаци (Synthetic data): Подаци који су моделирани да представљају статистичка својства оригиналних података; креирају се нове вредности података које, гледано у целини, чувају релевантна статистичка својства „стварног“ скупа података⁴⁸. Ово омогућава моделе обуке без приступања подацима из стварног света.

⁴⁷ The Royal Society. 2019 Protecting privacy in practice. See <https://royalsociety.org/topics-policy/projects/privacyenhancing-technologies/> (accessed 1 March 2024).

⁴⁸ *Ibid*

ПОГЛАВЉЕ 1

КАКО ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА ТРАНСФОРМИШЕ НАУЧНА ИСТРАЖИВАЊА

КАКО ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА ТРАНСФОРМИШЕ НАУЧНА ИСТРАЖИВАЊА

„Ослањамо се на [машинско учење] с циљем да би [наши експерименти] функционисали. Оно је постало уграђено у оно што радимо. Ако би избацили машинско учење из нашег процеса развоја, он би се распао.“
Учесник округлог стола Краљевског друштва

Анализа података великих размера и могућности препознавања образаца вештачке интелигенције (ВИ) представљају значајне могућности за унапређење научних истраживања и открића. Нови развој у машинском учењу, посебно, омогућава истраживачима да мапирају дефорестацију до појединачног дрвета⁴⁹; фармацеутским компанијама да развију нове терапије⁵⁰; и технолошким компанијама да открију нове материјале⁵¹. Овај развој представља нове могућности и изазове за природу и метод научног истраживања.

Ослањајући се на увиде са округлих столова, интервјуа и наручених истраживања, ово поглавље описује како ВИ мења научни подухват.

Вештачка интелигенција у науци: преглед

Наручена анализа употребе ВИ у науци (заснована на објављеној академској литератури и фокусирана на ширину, а не на дубину техника) показује да се примене ВИ налазе у свим STEM областима⁵². Постоји концентрација у одређеним областима као што су медицина; наука о материјалима; роботика; генетика; и, што није изненађујуће, рачунарство. Физичке науке и медицина изгледа доминирају када је у питању употреба технологија повезаних са ВИ. Најистакнутије технике ВИ у STEM областима укључују вештачке неуронске мреже (АНН); машинско учење (ML) (укључујући дубоко учење (DL)); обраду природног језика; и препознавање слика

У физичким наукама (нпр. физика, хемија, астрономија), вештачка интелигенција се примењује као метода за издвајање информација из брзо акумулираних података (нпр. подаци генерисани у Великом хадронском сударачу⁵³); за идентификацију образаца у веома великим скуповима података; и за моделирање физичких експеримената. У здравственим наукама (нпр. медицина, стоматологија, ветеринарске науке), користи се као техника за помоћ у откривању и предвиђању болести; за подршку клиничком доношењу одлука; и за унапређење хирургије, обуке и роботике. У биологији, користи се за анализу података са сензора; за подршку управљању усевиима и водама; и за предвиђање 3Д структура протеина.

Вештачка интелигенција и методе научног истраживања

Недавни развој вештачке интелигенције сугерише да би могло доћи до трансформативних промена у методама научног истраживања. Ове промене се усредсређују на повећање ефикасности у решавању постојећих задатака, промену процеса за генерисање знања или омогућавање нових

⁴⁹ Eyres A et al. 2024 LIFE: A metric for quantitatively mapping the impact of land-cover change on global extinctions. Cambridge Open Engage. (<https://doi.org/10.33774/coe-2023-gpn4p-v4>).

⁵⁰ Paul, D, Sanap, G, Shenoy, S, Kalyane, D, Kalia, K, Tekade, R. K. 2021 Artificial intelligence in drug discovery and development. *Drug discovery today*, **26**. 80–93. (<https://doi.org/10.1016/j.drudis.2020.10.010>)

⁵¹ Merchant, A, Batzner, S, Schoenholz, SS, Aykol, M, Cheon, G, Cubuk, ED. 2023 Scaling deep learning for materials discovery. *Nature*, **624**. 80–85. (<https://doi.org/10.1038/s41586-023-06735-9>)

⁵² Berman B, Chubb J, and Williams K, 2024. The use of artificial intelligence in science, technology, engineering, and medicine. The Royal Society. <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/> (accessed 7 May 2024)

⁵³ CERN. The Large Hadron Collider. See <https://home.cern/science/accelerators/large-hadron-collider> (accessed 22 April 2024)

механизма откривања. Следећи примери истраживачких активности предузетим за овај извештај су:

1. Растућа употреба дубоког учења у свим областима

Примена дубоког учења (ДУ) трансформише анализу података и генерисање знања. Његова употреба за аутоматско издвајање и учење карактеристика из сирових података, обраду опсежних скупова података и препознавање образаца ефикасно надмашује линеарне моделе засноване на машинском учењу⁵⁴. ДУ је нашло примену у различитим областима, укључујући здравствену заштиту, помоћ у откривању болести и откривању лекова, науци о клими, помажући у моделирању климатских образаца и откривању временских прилика. Значајан пример је примена ДУ од стране компаније Google DeepMind за развој AlphaFold-a, система за предвиђање савијања протеина који је решио 50 година стар изазов у биологији деценијама раније него што се очекивало⁵⁵.

Развој тачних и корисних модела заснованих на ДУ је изазован због своје природе црне кутије и варијација у проблемима и подацима из стварног окружења. Ово ограничава њихову експланаторну моћ и поузданост као научних алата⁵⁶. (Види поглавље 2)

2. Добијање увида из неструктурираних података

Главни изазов за истраживаче је коришћење неструктурираних података (подаци који не прате одређени формат или структуру, што отежава обраду, управљање и коришћење за проналажење образаца). Способност руковања неструктурираним подацима чини учење на даљину ефикасним за задатке који укључују препознавање слика и обраду природног језика (NLP). У здравству, на пример, подаци могу бити детаљни; мултимодални и фрагментирани⁵⁷⁵¹. Могу укључивати слике, процене текста или нумеричке вредности из процена и читавања. Сакупљачи података у систему здравствене заштите могу бележити ове податке у различитим форматима или помоћу различитог софтвера. Обједињавање ових података и њихово разумевање може помоћи истраживачима да направе предвиђања и моделирају потенцијалне здравствене интервенције. Слично томе, генеративни модели вештачке интелигенције могу допринети генерисању и конвертовању података у различите режиме и стандарде, који нису ограничени на врсту података унетих у алгоритам⁵⁸⁵².

„Имамо капацитет да забележимо много више [података] него раније. Живимо у поплави података. Зато се надамо да ће нам методе машинског учења помоћи да то схватимо, а затим да покренемо истинске, научне хипотезе.“

Учесник округлог стола Краљевског друштва

Друге технике као што је узрочно машинско учење (методе за процену узрока и последице у подацима)⁵⁹⁵³ могу помоћи у обради неструктурираних података учењем сложених нелинеарних

⁵⁴ Choudhary, A, Fox, G, Hey, T. 2023. Artificial intelligence for science: A deep learning revolution. *World Scientific Publishing Co. Pte Ltd.* (<https://doi.org/10.1142/13123>)

⁵⁵ Google DeepMind. AlphaFold. See: <https://deepmind.google/technologies/alphafold/> (accessed 5 March 2024)

⁵⁶ Sarker, I. H. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, **2**. 420. (<https://doi.org/10.1007/s42979-021-00815-1>)

⁵⁷ Healy, M. J. R. 1973. What Computers Can and Cannot Do. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **184**(1077), 375–378. (<https://doi.org/10.1098/rspb.1973.0056>)

⁵⁸ World Health Organization. 2024. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. See: <https://www.who.int/publications/i/item/9789240084759> (accessed 5 March 2024)

⁵⁹ Kaddour J, Lynch A, Liu Q, Kusner M J, Silva R. 2022. Causal machine learning: A survey and open problems. *arXiv preprint* (<https://doi.org/10.48550/arXiv.2206.15475>)

односа између променљивих^{60,61}54,55. Платформе које тврде да могу да стекну увиде из неструктурираних података укључују Benevolent AI⁶²56 (користећи неструктуриране податке из биомедицинске литературе) и Microsoft-ов Project Alexandria⁶³57 (фокусиран углавном на пословно знање).

3. Вишеструке симулације великих размера

Генеративна способност алата вештачке интелигенције да уче из постојећег садржаја и генеришу предвиђања новог садржаја пружа научницима могућност да покрећу тачна предвиђања. Генерисање симулација и синтетички подаци⁶⁴58 (вештачки генерисани подаци) или дигитални близанци⁶⁵59 (виртуелни прикази физичких инфраструктура) су примери како се алати засновани на вештачкој интелигенцији могу користити за обучавање система⁶⁶60. За молекуларна истраживања, ово подразумева коришћење дубоких неуронских мрежа које користе податке о томе како молекули интерагују како би прецизно симулирале понашање на атомском нивоу⁶⁷61. Употреба синтетичких података, посебно синтетичких података који чувају приватност, такође може помоћи у ублажавању изазова пристрасности података и заштити приватности појединаца⁶⁸62.

4. Убрзавање синтезе информација

Модел великих језика (LLM) и NLP технике се све више користе за убрзавање задатака заснованих на тексту као што су академско писање⁶⁹63, спровођење прегледа литературе или израда резимеа⁷⁰64.

⁶⁰ Sanchez P, Voisey J, Xia T, Watson H, O'Neil A, and Tsaftaris, S. 2022 Causal machine learning for healthcare and precision medicine. *R. Soc open sci.* **9**: 220638 (<https://doi.org/10.1098/rsos.220638>)

⁶¹ Royal Society roundtable on large language models, July 2023

⁶² Benevolent AI. 2019 Extracting existing facts without requiring any training data or hand-crafted rules. See <https://www.benevolent.com/news-and-media/blog-and-videos/extracting-existing-facts-without-requiring-any-training-data-or-hand-crafted-rules/> (accessed 21 December 2023)

⁶³ Rajput S, Winn J, Moneypenny N, Zaykov Y, and Tan C. 2021 Alexandria in Microsoft Viva Topics: from big data to big knowledge. 26 April 2021. See <https://www.microsoft.com/en-us/research/blog/alexandria-in-microsoft-viva-topics-from-big-data-to-big-knowledge/> (accessed 21 December 2023)

⁶⁴ Jordon *et al.* 2023 Synthetic Data – what, why and how? See <https://royalsociety.org/news-resources/projects/privacy-enhancing-technologies/> (accessed 21 December 2023)

⁶⁵ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023)

⁶⁶ Jordon *et al.* 2023 Synthetic Data – what, why and how? See <https://royalsociety.org/news-resources/projects/privacy-enhancing-technologies/> (accessed 21 December 2023).

⁶⁷ Zhang L, Han J, Wang H, Car R, Weinan E. 2018 Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys Rev Lett.* 2018 Apr 6;120(14):143001. (<https://doi.org/10.1103/PhysRevLett.120.143001>). PMID: 29694129

⁶⁸ The Royal Society. 2023 From privacy to partnership. See <https://royalsociety.org/topics-policy/projects/privacyenhancing-technologies/> (accessed 21 December 2023).

⁶⁹ Lin Z. 2023 Why and how to embrace AI such as ChatGPT in your academic life. *R. Soc. Open Sci.* **10**: 230658 (<https://doi.org/10.1098/rsos.230658>)

⁷⁰ Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus.* 2023 Feb 19;15(2):e35179. (<https://doi.org/10.7759/cureus.35179>)

Примери аутоматизованих алата за преглед литературе укључују Semantic Scholar⁷¹65, Elicit⁷²66 и Consensus⁷³67. Такође је доступан на истакнутим платформама као што су GPT4 и Gemini. Корисни случајеви употребе укључују коришћење LLM-а за побољшање квалитета академског писања, помоћ у преводу или емулацију специфичних стилова писања (нпр. израда лаичких резимеа). Поред академских текстова, могу се користити и за поједностављивање административних задатака и помоћ у изради пријава за грантове. Ови алати би такође могли побољшати приступачност за истраживаче из различитих средина (нпр. особе које не говоре енглески језик и неуродивергентне особе) који конзумирају и производе академски садржај на више језика и формата⁷⁴68. Ови алати такође имају ограничења, укључујући потенцијал да погоршају пристрасности из података о обуци (нпр. пристрасност ка позитивним резултатима⁷⁵69, језичке пристрасности⁷⁶70 или географске пристрасности⁷⁷71), нетачности и непоуздане научне улазе⁷⁸72. Као алат за писање, они такође имају ограничену способност да схвате нијансиране вредносне судове, помогну у стварању научног значења⁷⁹73 или артикулишу сложеност научног истраживања⁸⁰74. Такође постоји забринутост да употреба мастер студија за академско писање ризикује смањење креативних и интердисциплинарних аспеката научних открића⁸¹75. Поред тога, постоје питања о утицају мастер студија на интелектуалну својину (ИС).

5. Решавање сложених изазова кодирања

Развој софтверског кода за рачунарску анализу постао је важан аспект савременог научног подухвата. На пример, мастер студије за рачунарску анализу – које су дизајниране да анализирају текстуалне уносе и генеришу одговоре за које утврде да су вероватно тачни – могу се користити за генерисање софтверског кода у различитим језицима кодирања. Ово пружа прилику научним истраживачима да конвертују код из једног рачунарског језика у други или између апликација⁸²76. Чак и ако резултат није тачан у првом покушају, ови модели се могу користити као асистенти у кодирању како би се идентификовале грешке у кодирању, давали предлози и уштедело време.

⁷¹ Semantic Scholar: AI-powered research tool. See <https://www.semanticscholar.org/> (accessed 21 December 2023)

⁷² Elicit: The AI research assistant. See <https://elicit.com/> (accessed 21 December 2023)

⁷³ Consensus: AI search engine for research. See <https://consensus.app/> (accessed 21 December 2023)

⁷⁴ Royal Society and Department for Science, Innovation, and Technology workshop on horizon scanning AI safety risks across scientific disciplines, 2023.

⁷⁵ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/>. (accessed 7 May 2024)

⁷⁶ Barrot JS, 2023. Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57.100745

⁷⁷ Skopec M, Issa H, Reed J, Harris M. 2020. The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis. *Research integrity and peer review*, 5. 1-14. (<https://doi.org/10.1186/s41073-019-0088-0>.)

⁷⁸ Sanderson K. 2023. GPT-4 is here: what scientists think. *Nature*, 615.773. 30 March 2023. See <https://www.nature.com/articles/d41586-023-00816-5.pdf> (accessed 21 December 2023)

⁷⁹ Birhane A, Kasirzadeh A, Leslie D, Wachter S. 2023. Science in the age of large language models. *Nature Reviews Physics*, 1-4 (<https://doi.org/10.1038/s42254-023-00581-4>)

⁸⁰ Bender E, Koller A. 2020 Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 5185–5198

⁸¹ Royal Society roundtable on large language models, July 2023

⁸² Royal Society roundtable on large language models, July 2023

Истакнути примери укључују Мајкрософтов Copilot⁸³77; OpenAI-ов GPT4⁸⁴78; Meta-ин Code Llama⁸⁵79; и Google DeepMind-ов Gemini.⁸⁶80

6. Аутоматизација задатака

Алати вештачке интелигенције могу аутоматизовати низ задатака који захтевају много времена и рада у оквиру процеса научног рада.⁸⁷81 Аутоматизација може довести до повећања продуктивности за научнике⁸⁸82 и коришћење потенцијал за тестирање различитих хипотеза које превазилазе људске могућности. Google DeepMind је, 2023. године, навео два таква примера: FunSearch⁸⁹83 и GNoME⁹⁰84. Употреба роботских истраживачких асистената такође доприноси аутоматизацији лабораторијских процеса рада (видети студију случаја 2). Робот који је развио Универзитет Абериствит, 2009. године, постао је прва машина која је самостално открила нова научна знања⁹¹85. Робот је програмиран да самостално дизајнира експерименте, бележи и процењује резултате и поставља нова питања – аутоматизујући цео процес истраживања⁹²86. Научници у области роботике, надовезујући се на овај пробој, дају доприносе који омогућују убрзавање процеса откривања, а истовремено и смањују трошкове, неизвесност и људске грешке у лабораторијама⁹³87. Како истраживање постаје све аутоматизованије, постоји забринутост да би будуће генерације научника могле постати неквалификоване у основним вештинама као што су генерисање хипотеза, експериментални дизајн и контекстуална интерпретација⁹⁴88. Методолошка транспарентност и разумевање узрочно-последичних веза такође би могли да опадну, а претерано

⁸³ GitHub. Copilot – Your AI pair programmer. See <https://github.com/features/copilot> (accessed 21 December 2023).

⁸⁴ Open AI. GPT4. See <https://openai.com/gpt-4> (accessed 21 December 2023).

⁸⁵ Meta. 2023 Introducing Code Llama, a state-of-the-art large language model for coding. *Meta*. 24 August 2023. See <https://ai.meta.com/blog/code-llama-large-language-model-coding/> (accessed 21 December 2023)

⁸⁶ Google DeepMind. Gemini. See <https://deepmind.google/technologies/gemini/#introduction> (accessed 21 December 2023)

⁸⁷ Xie, Y, Sattari, K, Zhang, C, & Lin, J. 2023 Toward autonomous laboratories: Convergence of artificial intelligence and experimental automation. *Progress in Materials Science*, **132**. 101043. (<https://doi.org/10.1016/j.pmatsci.2022.101043>)

⁸⁸ OECD. 2023. Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, *OECD Publishing, Paris* (<https://doi.org/10.1787/a8d820bd-en>).

⁸⁹ Fawzi A and Paredes B. 2023. FunSearch: Making new discoveries in mathematical sciences using Large Language Models. *Google DeepMind*. See <https://deepmind.google/discover/blog/funsearch-making-new-discoveries-in-mathematical-sciences-using-large-language-models/> (accessed 21 December 2023).

⁹⁰ Merchant A and Cubuk E. 2023 Millions of new materials discovered with deep learning. *Google DeepMind*. See <https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/> (accessed 21 December 2023).

⁹¹ University of Cambridge. Robot scientist becomes first machine to discover new scientific knowledge. See: <https://www.cam.ac.uk/research/news/robot-scientist-becomes-first-machine-to-discover-new-scientific-knowledge> (accessed 3 March 2024)

⁹² Sparkes A *et al.* 2010. Towards Robot Scientists for autonomous scientific discovery. *Autom Exp 2*, 1 (<https://doi.org/10.1186/1759-4499-2-1>)

⁹³ University of Cambridge. Artificially-intelligent Robot Scientist ‘Eve’ could boost search for new drugs. See: <https://www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs> (accessed 7 March 2024)

⁹⁴ Lin Z. 2023 Why and how to embrace AI such as ChatGPT in your academic life. *R Soc Open Sci*. 2023 Aug 23;10(8):230658 (<https://doi.org/10.1098/rsos.230658>.)

наглашавање рачунарских техника ризикује да одврати научнике који траже креативне изразе у свом раду⁹⁵89

Вештачка интелигенција и природа научног истраживања

Поред утицаја вештачке интелигенције на методе научног истраживања, постоји потенцијално трансформативан утицај на саму природу научног подухвата. Ови утицаји се првенствено односе на распрострањеност истраживања заснованих на великим базама података, ослањање на рачунарску снагу и нове начине организовања вештина и рада у научном процесу.

На основу активности предузетих за овај извештај, издвојено је следећих шест тема које се сматрају кључним за утицај вештачке интелигенције на природу научног истраживања.

1. Рачунари и људски рад као темељне инфраструктуре вештачке интелигенције

Скуп дигиталних инфраструктура и људског рада је основа главних примена вештачке интелигенције⁹⁶90. Дигитална инфраструктура се односи на уређаје који прикупљају податке, личне рачунаре на којима се анализирају и суперрачунаре који покрећу анализу података великих размера. Људски рад се односи на чин прикупљања, чишћења и означавања података, као и на чин пројектовања, тестирања и имплементације. Врсте потребне дигиталне инфраструктуре укључују суперрачунаре (нпр. оне укључене у HPC-UK⁹⁷91 и EuroHPC JU⁹⁸92); технологије за побољшање поштовања приватности;⁹⁹93 и објекте за складиштење података (нпр. центри података) Решења заснована на облаку, која не захтевају од корисника да поседују физичку инфраструктуру (нпр. за складиштење података), укључују Amazon Web Services¹⁰⁰94 и Oracle Cloud Infrastructure.¹⁰¹95

2. Доминација истраживања усмереног на велике базе податка

Могућност прикупљања велике количине података (великих и хетерогених облика података који су прикупљени без строгог експерименталног дизајна¹⁰²96) и комбиновања ових са другим скуповима података представља јасне и значајне могућности за научни подухват. Значајни резултати добијени применом вештачке интелигенције на ове скупове података већ су допринели у низу примера позитивних примена, од ублажавања утицаја COVID-19 до борбе против климатских промена (видети студију случаја 3)¹⁰³97. Све ово ће вероватно наставити да преобликује истраживачке подухвате ка већој усмерености на вештачку интелигенцију и велике базе

⁹⁵ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/>. (accessed 7 May 2024)

⁹⁶ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/> (accessed 7 May 2024)

⁹⁷ HPC-UK. UK HPC Facilities. See <https://www.hpc-uk.ac.uk/facilities/> (accessed 21 December 2023)

⁹⁸ The European High Performance Computing Joint Undertaking. See https://eurohpc-ju.europa.eu/index_en (accessed 21 December 2023).

⁹⁹ The Royal Society. Privacy Enhancing Technologies. See <https://royalsociety.org/topics-policy/projects/privacyenhancing-technologies/> (accessed 21 December 2023).

¹⁰⁰ Amazon Web Services. Cloud Computing Services. See <https://aws.amazon.com/> (accessed 21 December 2023)

¹⁰¹ Oracle. Cloud Infrastructure. See <https://www.oracle.com/cloud/> (accessed 21 December 2023).

¹⁰² The Royal Society. 2017 Machine Learning: The power and promise of computers that learn by example. See <https://royalsociety.org/topics-policy/projects/machine-learning/> (accessed 21 December 2023)

¹⁰³ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023)

податка¹⁰⁴98. Међутим, могућност укључивања у истраживање усмерено на велике базе податка остаје зависна од приступа рачунарској инфраструктури која омогућава обраду великих хетерогених скупова података.

Доминација истраживања усмереног на велике базе податка такође има импликације на истраживања у којима су доступни само непотпуни или малобројни подаци. Без пажљивог управљања, ризикује се смањење улагања у истраживање и подршке приоритетним областима (нпр. предметима или регионима) где је прикупљање примарних података у значајној размери ограничено, отежано или није пожељно. Такође је вероватно да ће се повећати заинтересованост за технике које имају за циљ проширивање података и употребу синтетичких података. Случај истраживања ретких болести (видети студију случаја 1) илуструје примену вештачке интелигенције у истраживању базираном на малом броју података.

3. Отворена наспрам затворене науке

Отворена наука, која тежи да отвори цео процес истраживања и објављивања (укључујући, али не ограничавајући се на отворене податке; отворене протоколе; отворени код; и транспарентну рецензију), је принцип и пракса коју заговара Краљевско друштво и други¹⁰⁵99. Такође је промовишу велике технолошке компаније, укључујући Мета и OpenAI, иако је ово оспоравано као „аспиративно“ или чак „маркетиншко“ пре него технички дескриптор¹⁰⁶100. Поред тога што обезбеђују транспарентност, приступи отворене науке могу омогућити репликацију експеримената, шири јавни увид у истраживачке производе¹⁰⁷101 и даље право свих да учествују у научном напретку¹⁰⁸102. Међутим, све већа употреба власничке вештачке интелигенције представља изазове за отворену науку. Истраживачи се све више ослањају на алате које развијају и одржавају приватне компаније (видети поглавље 4), иако унутрашње функционисање може остати непрозирно¹⁰⁹103. Ово је погоршано непрозирношћу података за обуку који су основа истакнутих алата вештачке интелигенције. Лоша транспарентност ризикује ограничавање корисности алата вештачке интелигенције за решавање проблема из стварног света, јер креатори политике и научници можда неће сматрати резултате генерисане вештачком интелигенцијом поузданим за важне одлуке¹¹⁰104. Такође поткопава напоре да се открију и испитају негативни утицаји или дискриминаторни ефекти¹¹¹105.

¹⁰⁴ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/>. (accessed 7 May 2024)

¹⁰⁵ The Royal Society. Open Science. See <https://royalsociety.org/journals/open-access/open-science/> (accessed 21 December 2023)

¹⁰⁶ Widder D, West S, Whittaker M. 2023 Open (for business): Big tech, concentrated power, and the political economy of Open AI. *SSRN*. (<https://doi.org/10.2139/ssrn.4543807>)

¹⁰⁷ The Royal Society. 2022 The online information environment. See <https://royalsociety.org/topics-policy/projects/online-information-environment/> (accessed 21 December 2023).

¹⁰⁸ UNESCO Recommendation on Open Science. 2021. See: <https://www.unesco.org/en/legal-affairs/recommendationopen-science> (accessed 6 February 2024)

¹⁰⁹ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/> (accessed 7 May 2024)

¹¹⁰ UNESCO Recommendation on Open Science. 2021. See: <https://www.unesco.org/en/legal-affairs/recommendationopen-science> (accessed 6 February 2024)

¹¹¹ UNESCO Recommendation on ethics of artificial intelligence. 2022. See: <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> (accessed 6 February 2024)

Потпуно отворени приступ који подстиче објављивање скупова података и модела без заштитних ограда или смерница такође можда није пожељан, јер скупове података или моделе могу манипулисати лоши актери¹¹²106. Потребни су контекстуално специфични и са вештачком интелигенцијом компатибилни приступи отворене науке како би се побољшао надзор и транспарентност^{113,114}107,108

4. Оспоравање појмова транспарентности и објашњивости

Научни метод се генерално може описати као чин стварања хипотезе, спровођења експеримента, бележења његовог исхода и усавршавања оригиналне хипотезе у складу са резултатима. Овај концепт датира више од хиљаду година од Хасана Ибн ел-Хајтама, који је наглашавао потребу за експерименталним подацима и репродуктивношћу резултата.¹¹⁵109 У основи овог и других приступа научној методологији је „потрага за објашњењима као основни циљ науке“¹¹⁶110. Ово се доводи у питање због недавних развоја вештачке интелигенције због нелинеарних односа који се могу извести из великих база података и опште природе црне кутије алата вештачке интелигенције^{117,118}111,112. Ова открића би могла бити трансформативна за друштво. Међутим, ако се истраживачи претерано ослоне¹¹⁹113 на вештачку интелигенцију за тумачење и анализу резултата, можда неће моћи да објасне како су дошли до закључака или да пруже информације потребне за репродукцију студије.¹²⁰114 Ово не би испунило праг како се традиционално прихвата да се спроводи наука, јер би се колеге мучиле да потврде или оповргну хипотезу. (Видети Поглавље 2 за више детаља о изазовима које вештачка интелигенција представља у погледу објашњивости и репродуктивности).

5. Наука као интердисциплинарни подухват

Успешна примена вештачке интелигенције у научним истраживањима и њено превођење у вредност у стварном свету захтева интердисциплинарне вештине и разумевање. Компјутерски научници који желе да примене вештачку интелигенцију за решавање великих научних проблема морају бити у стању да процене моделе вештачке интелигенције у другим истраживачким областима (нпр. здравство, климатска наука). Слично томе, научници који нису стручњаци за

¹¹² Vincent J. 2023 OpenAI co-founder on company's past approach to openly sharing research: 'We were wrong'. The Verge. 15 March 2023. See <https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closedresearch-ilya-sutskever-interview> (accessed 21 December 2023)

¹¹³ House of Lords. 2024 Large language models and generative AI. Report of Session 2023-24. See: <https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm> (accessed 2 February 2024)

¹¹⁴ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

¹¹⁵ Al-Khalili J. 2009 The 'first true scientist'. *BBC News*. 4 January 2009. See <http://news.bbc.co.uk/1/hi/sci/tech/7810846.stm> (accessed 21 December 2023).

¹¹⁶ Maxwell N. 1972 A Critique of Popper's Views on Scientific Method. *Philosophy of Science*, 39(2), 131-152. (doi:10.1086/288429)

¹¹⁷ Succi S, Coveney P. 2019 Big data: The end of the scientific method? *Phil. Trans. R. Soc. A*. 377: 20180145. (<https://doi.org/10.1098/rsta.2018.0145>)

¹¹⁸ The Royal Society. 2019 Explainable AI: the basics. See <https://royalsociety.org/topics-policy/projects/explainable-ai/> (accessed 21 December 2023).

¹¹⁹ Bućinca Z, Malaya M, Gajos K. 2021 To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5:188. (<https://doi.org/10.1145/3449287>)

¹²⁰ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/>. (accessed 7 May 2024)

вештачку интелигенцију морају да разумеју како да ефикасно користе алате и технике вештачке интелигенције за своје експерименте. Интегрисање знања из различитих области и система знања такође може довести до тачнијих модела и неговати истраживање вођено радозналошћу (изван комерцијално вођених интересовања)¹²¹115. Иницијатива eScience у Великој Британији (2001–2008)^{122,123}116,117 истиче се као напор да се негује интердисциплинарна сарадња неговањем културе у којој научници и стручњаци за рачунарске науке раде заједно. Текуће иницијативе попут Института Алана Тјуринга¹²⁴118 и Школе за одрживост Државног универзитета Аризоне¹²⁵119 такође настављају да се залажу за интердисциплинарне приступе. Међутим, интердисциплинарност је угушена изолованим институцијама и недовољним могућностима за напредовање у каријери. Интердисциплинарност не мора бити ограничена само на природне науке, већ се може стећи вредност радом научника са истраживачима у уметности, хуманистичким и друштвеним наукама. Пример за то укључује значај уметника у дизајну корисничког искуства импресивних окружења^{126,127}120,121 (Видети поглавље 3 за више детаља о интердисциплинарности у истраживањима заснованим на вештачкој интелигенцији).

6. Спој људске стручности са аутоматизацијом вештачке интелигенције

Окретање ка аутоматизацији нуди могућности за комбиновање људске стручности са ефикасношћу коју омогућава вештачка интелигенција. Вештачка интелигенција се може користити или као допуна раду научнику тако што помаже или повећава људске способности; или да се развију аутономни механизми за откривање (видети слику 1)¹²⁸122. У овом спектру, човек научник остаје неопходан за контекстуално научно разумевање. Растућа употреба алата вештачке интелигенције такође ризикује да научнике учини рањивим на „илузије разумевања“ у којима је само ограничен скуп гледишта и метода представљен у резултатима¹²⁹123. Постоји потреба за даљим разумевањем приступа „човек-укључен“ (“human-in-the-loop”) који препознају вештачку интелигенцију као допуну људском расуђивању и улогу људске интервенције како би се осигурао квалитет резултата.

Опис три могућа начина на које вештачка интелигенција може допринети научном разумевању дат је слици 1, страна 31 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

¹²¹ The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023

¹²² Hey T, Trefethen A. 2002 The UK e-Science Core Programme and the Grid Hey, T., & Trefethen, A. E. International Conference on Computational Science (pp. 3-21). Berlin, Heidelberg: Springer Berlin Heidelberg ([https://doi.org/10.1016/S0167-739X\(02\)00082-1](https://doi.org/10.1016/S0167-739X(02)00082-1))

¹²³ Hey T. 2005. e-Science and open access. See https://www.researchgate.net/publication/28803295_E-Science_and_Open_Access (accessed 7 May 2024)

¹²⁴ The Alan Turing Institute. Research. See <https://www.turing.ac.uk/research> (accessed 21 December 2023).

¹²⁵ Arizona State University - School of Sustainability. See <https://schoolofsustainability.asu.edu/> (accessed 21 December 2023)

¹²⁶ The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/> (accessed 21 December 2023).

¹²⁷ *Ibid.*

¹²⁸ Krenn M. et al. 2022. On scientific understanding with artificial intelligence. *Nature Reviews Physics* **4**. (<https://doi.org/10.1038/s42254-022-00518-3>)

¹²⁹ Messeri L, Crockett MJ. 2024 Artificial intelligence and illusions of understanding in scientific research. *Nature*. Mar;627(8002):49-58. (<https://doi.org/10.1038/s41586-024-07146-0>.)

А детаљан опис је дат у дијаграм који описује три могућа начина на које вештачка интелигенција може допринети научном разумевању. „Рачунарски микроскоп“ се односи на улогу вештачке интелигенције у пружању информација путем напредне симулације и представљања података који се не могу добити експериментисањем. „Извор инспирације“ се односи на сценарије у којима вештачка интелигенција пружа информације које проширују обим људске маште или креативности. „Агент разумевања“ илуструје сценарио у којем аутономни системи вештачке интелигенције могу да деле увиде са стручњацима претварајући запажања у нова знања. За сада нема доказа који би указивали на то да рачунари могу деловати као прави агенти научног разумевања. Видети реф. 124. Krenn M. et al. 2022. On Scientific Understanding with Artificial Intelligence.

„Сви желе блиставу фонтану, али врло мало људи размишља о досадном водоводном систему испод ње.“

Pete Buttigieg Participant in the US-UK Scientific Forum on Researcher Access to Data

ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА И ПРИСТУП ВИСОКОКВАЛИТЕТНИМ ПОДАЦИМА

Подаци су основа система вештачке интелигенције. Израз и принцип „смеће унутра, смеће напоље“, који датира из раних дана рачунарства¹³⁰125 и списка Чарлса Бебиџа FRS¹³¹126, остају важећи и данас. Подаци лошег квалитета који су непотпуни, нетачни или нерепрезентативни могу довести до обманујућих резултата. Обезбеђивање висококвалитетних скупова података за обуку система вештачке интелигенције подразумева решавање изазова као што су поверење, приступ, пристрасност, доступност и интероперабилност током животног циклуса података. Ослањајући се на Америчко-британски научни форум о приступу истраживача подацима из 2023. године, који су организовали Краљевско друштво и Национална академија наука САД¹³²127; и интервјуе, следеће теме су се појавиле као кључни изазови повезани са коришћењем података за научна истраживања заснована на вештачкој интелигенцији

1. Запремина и хетерогеност

Подаци прикупљени за научне експерименте могу бити изузетно велики, мерећи се у терабајтима, петабајтима и ексабајтима¹³³128. Овај изазов запремине примењује се на различите области, укључујући геномику, физику високих енергија, климатску науку и астрономију. Пример представљен на Научном форуму САД-УК¹³⁴129 је телескоп „Event Horizon“ (Технолошки институт Џорџије) који је направио прве две фотографије црних рупа.¹³⁵130 Пројекат, у којем је учествовало десет телескопа широм света, забележио је један петабајт података по ноћи. Огромна

¹³⁰ Mellin W. 1957 Work with new electronic ‘brains’ opens field for army math experts. *The Hammond Times*.

¹³¹ Babbage C. 1964. *Passages from the life of a philosopher*. Cambridge, UK: Cambridge University Press.

¹³² National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>

¹³³ Today, a typical PC or laptop comes with one terabyte storage. Petabytes are akin to the storage capacity of a thousand of these PCs and exabytes are akin to the storage capacity of a million

¹³⁴ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>

¹³⁵ Event Horizon Telescope. See <https://eventhorizontelescope.org/> (accessed 21 December 2023)

величина ове врсте података и њихова често хетерогена природа отежавају постизање циљева као што је интероперабилност. Овај изазов захтева интегрисана и централна спремишта која пружају дугорочно управљање и приступ истраживачима, дисеминацију и анализу у скоро реалном времену, као и решења за анотацију хетерогених података за обуку. На пример, Национална управа за океане и атмосферу експериментира са решењем за видео игру под називом FathomVerse, тражећи од играча да идентификују широке категорије врста које могу да виде на видео снимцима са истраживачких бродова¹³⁶131 Успостављање великих инфраструктура података захтева даље разматрање њиховог одржавања и утицаја на животну средину¹³⁷132

2. Улога институција за базе податка

Институције за податке (нпр. архиве, статистичке агенције, складишта података) могу играти важну улогу у олакшавању приступа истраживача подацима. Ове институције имају низ функција, укључујући:

- Заштиту осетљивих података и одобравање приступа под ограниченим условима.
- Комбиновање или повезивање података из више извора и пружање увида и других услуга онима који су допринели подацима.
- Креирање отворених скупова података којима свако може приступити, користити и делити ради унапређења одређене мисије или циља.
- Деловање као чувар података које држе друге организације.
- Развој и одржавање идентификатора, стандарда и друге инфраструктуре за сектор или област, као што је регистрација идентификатора или објављивање отворених стандарда.
- Омогућавање људима да преузму активнију улогу у управљању подацима о себи и својим заједницама.

Њихова фундаментална улога у одржавању квалитета података доступних за научна истраживања чини их сличним основној инфраструктури за алате вештачке интелигенције. Као такви, они коштају новац и захтевају одржавање.¹³⁸133 Примери институција за податке укључују веб-сајт отворених података Владе САД, Data.gov¹³⁹134, Међууниверзитетски конзорцијум за политичка и друштвена истраживања Универзитета у Мичигену (који садржи податке о више од 19.000 студија друштвених и бихевиоралних наука)¹⁴⁰135 и Канцеларију за националну статистику Уједињеног Краљевства¹⁴¹136.

3. Дељење осетљивих података

Ограничења дељења осетљивих података могу блокирати потенцијалне научне продоре. У области здравствене заштите, на пример, дељење и обрада здравствених података је сложено због своје поверљиве и фрагментиране природе (како унутар институција, тако и преко граница). Иако напредак у медицинском снимању, тексту, звуку и вештачкој интелигенцији нуди нове могућности за дијагнозу и лечење, он такође ризикује откривање осетљивих информација о пацијентима путем снимања и метаподатака. Слично томе, лоши актери такође могу да искористе податке о животној средини (нпр. податке о падавинама, крчењу шума или криволову) као оружје како би изазвали

¹³⁶ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>.

¹³⁷ SKAO. Portuguese prove SKA green energy system. See: <https://www.skao.int/en/impact/440/portuguese-proveska-green-energy-system> (accessed 21 March 2024)

¹³⁸ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>

¹³⁹ Data.gov. See <https://data.gov/> (accessed 21 December 2023).

¹⁴⁰ ICPSR. See <https://www.icpsr.umich.edu/web/pages/index.html> (accessed 21 December 2023).

¹⁴¹ Office for National Statistics. See <https://www.ons.gov.uk/> (accessed 21 December 2023).

претње по националну безбедност и животну средину. Истраживачи позивају на дефиницију „осетљивих података“, која разматра како подаци који су предмет експлоатације, злоупотребе и погрешног тумачења¹⁴²137 могу проузроковати друштвену и еколошку штету¹⁴³138

Употреба поузданих истраживачких окружења и технологија за побољшање приватности (укључујући приступе засноване на вештачкој интелигенцији, као што је федеративно машинско учење), омогућава истраживачима да моделирају проблеме без потребе за приступом подацима, нудећи потенцијално техничко решење за решавање проблема у вези са осетљивим подацима. Ово је детаљно објашњено у извештају Краљевског друштва из 2019. године „Заштита приватности у пракси“¹⁴⁴139 и извештају из 2023. године „Од приватности до партнерства“ (који садржи различите случајеве употребе).¹⁴⁵140 Јавно поверење и прихvatљивост у вези са коришћењем осетљивих скупова података који се односе на људе (нпр. здравствене информације, демографија, локација итд.) такође су неопходни. Као што је наведено у извештају Краљевског друштва из 2023. године „Стварање отпорних и поузданих система података“, поверење у дељење података захтева јасноћу сврхе и транспарентност у токовима података, као и робусне системе за безбедност и приватност¹⁴⁶141. Актери из приватног сектора, као што су IBM, Microsoft и Siemens, баве се забринутостима јавности успостављањем заједница поверења¹⁴⁷142. Други приступи укључују оквири за управљање подацима који подстичу јавност да се укључи у научне пројекте засноване на подацима, уз задржавање контроле над својим подацима (нпр. акције донирања података¹⁴⁸143).

Студија случаја 1

ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА И ДИЈАГНОЗЕ РЕТКИХ БОЛЕСТИ

Ретка болест је стање које погађа мање од једног од 2.000 људи и често је окарактерисана разноврсним, сложеним и преклапајућим генетским манифестацијама¹⁴⁹144. За више од 7.000 ретких болести, описаних широм света, само 5% има лечење¹⁵⁰145. Недостатак разумевања основних узрока, фрагментирани подаци о пацијентима и неадекватне политике допринели су томе да дијагноза и лечење ретких болести постану изазов за јавно здравље¹⁵¹146.

¹⁴² Leonelli S, Williamson H. 2022 Introduction: Towards Responsible Plant Data Linkage. In: Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development. Springer International Publishing (https://doi.org/10.1007/978-3-031-13276-6_1).

¹⁴³ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>

¹⁴⁴ The Royal Society. 2019 Protecting privacy in practice. See <https://royalsociety.org/topics-policy/projects/privacyenhancing-technologies/> (accessed 21 December 2023).

¹⁴⁵ *Ibid*

¹⁴⁶ The Royal Society. 2023 Creating resilient and trusted data systems. See <https://royalsociety.org/topics-policy/projects/data-for-emergencies/> (accessed 21 December 2023)

¹⁴⁷ Charter of Trust. See: www.charteroftrust.com (accessed 21 December 2023)

¹⁴⁸ The Tidepool Big Data Donation Project. See: <https://www.tidepool.org/bigdata> (accessed 21 December 2023)

¹⁴⁹ Department of Health and Social Care. 2021. The UK Rare Diseases Framework. See: <https://www.gov.uk/government/publications/uk-rare-diseases-framework/the-uk-rare-diseases-framework> (accessed 30 September 2023).

¹⁵⁰ Brasil S, Pascoal C, Francisco R, Dos Reis Ferreira V, Videira PA, Valadão AG. 2019 Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter? (<https://doi.org/10.978 10.3390/genes10120978>)

¹⁵¹ Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. 2021 Opportunities and challenges for machine learning in rare diseases. *Frontiers in Medicine*, 8, 747612. (<https://doi.org/10.3389/fmed.2021.747612>)

Примена техника машинског учења и генеративне вештачке интелигенције нуди прилику за превазилажење неких од ових ограничења. Истраживачи ретких болести користе технике машинског учења за анализу високодимензионалних (high-dimensional datasets) скупова података, као што су високодимензионални молекуларни подаци, како би идентификовали релевантне биомаркере за познате болести или идентификовали нове болести¹⁵²147. Прелазак на дигитализацију здравствених картона такође ствара могућности за бржу идентификацију пацијената са ретким болестима. Обећавајуће примене показују потенцијал за побољшање ниских стопа дијагностике, третмана и процеса развоја лекова¹⁵³148

Примене вештачке интелигенције у области ретких болести

• Коришћење медицинског снимања за рандијагнозу:

Клиничари користе вештачку интелигенцију да би пронашли обрасце у великим скуповима података о пацијентима, укључујући генетске податке и клиничке картотеке, који могу указивати на присуство ретке болести. Машинско учење је посебно корисно за анализу мултимодалних података из различитих извора, укључујући податке снимања (нпр. MRI, рендгенске снимке) што постаје стандардна пракса за разумевање манифестације болести¹⁵⁴149. На пример, истраживачи на Институту за геномску статистику и биоинформатику на Универзитету у Бону користе дубоке неуронске мреже (DNN) и рачунарску анализу лица како би убрзали дијагнозу ултра-ретких и нових поремећаја¹⁵⁵150.

• Побољшање могућности за аутоматизовану дијагнозу:

Технике машинског учења могу се користити и за побољшање аутоматизоване дијагностичке подршке за клиничаре. Примена машинског учења на веома велике мултимодалне скупове здравствених података, као што је, на пример, UK Biobank¹⁵⁶151, ствара нове могућности за откривање непознатих и нових варијанти које могу допринети молекуларној дијагнози ретких болести¹⁵⁷152. Слично томе, машинско учење се примењује да би се утврдило да ли се подаци у здравственим скуповима података могу користити за идентификацију пацијената који нису претходно тестирани на ретке болести и који можда нису дијагностиковани^{158,159}153,154.

• Убрзати лечење и откривање лекова:

Модел машинског учења, посебно генеративна вештачка интелигенција, могу се искористити за убрзање процеса откривања лекова. Модел прегледају молекуларне библиотеке како би предвидели потенцијалне кандидате за лекове и проценили њихову ефикасност у лечењу

¹⁵² Banerjee J *et al.* 2023 Machine learning in rare disease. *Nat Methods* 20, 803–814. (<https://doi.org/10.1038/s41592-023-01886-z>)

¹⁵³ Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. 2020 The use of machine learning in rare diseases: a scoping review. *Orphanet J Rare Dis.* (<https://doi.org/10.1186/s13023-020-01424-6>)

¹⁵⁴ *Ibid.*

¹⁵⁵ Hsieh TC, Krawitz PM. 2023 Computational facial analysis for rare Mendelian disorders. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics.* (<https://doi.org/10.1002/ajmg.c.32061>)

¹⁵⁶ UK Biobank. See: <https://www.ukbiobank.ac.uk/> (accessed 21 December 2023)

¹⁵⁷ Turro E *et al.* 2020 Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* 583, 96–102 (<https://doi.org/10.1038/s41586-020-2434-2>)

¹⁵⁸ Cohen A M *et al.* Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria. *PLoS.* (<https://doi.org/10.1371/journal.pone.0238277>)

¹⁵⁹ Hersh W R *et al.* 2022 Clinical study applying machine learning to detect a rare disease: results and lessons learned. *JAMIA Open*, 5. (<https://doi.org/10.1093/jamiaopen/ooac053>)

одређених ретких болести.¹⁶⁰155 Овом облашћу обично доминирају фармацеутске компаније приватног сектора као што су Insilico Medicine, Recursion или Healx¹⁶¹156.

Изазови у вези података за примену вештачке интелигенције за студије ретких болести

- **Ограничена доступност података:** Ретке болести погађају веома мали проценат светске популације. Релевантни подаци – ако су доступни – су изоловани, расути, иза платних зидова или у комерцијалном власништву. Ова оскудица података може отежати обуку тачних и робусних модела вештачке интелигенције¹⁶²157. Ово је погоршано недостатком канала за координацију између лабораторија и институција ради интеграције и укрштања скупова података.
- **Пристрасни и нерепрезентативни скупови података:** Када су подаци о ретким болестима доступни, они могу бити нерепрезентативни, стварајући проблеме који се крећу од потенцијално лажно позитивних или негативних резултата до недовољне заступљености различитих старосних група или етничких мањина. Неуравнотежени подаци могу довести до пристрасних и „преоптерећених“ модела који се ослањају на обрасце који су јединствени за податке за обуку и стога лоше функционишу на новим скуповима података, са импликацијама на преносивост и генерализацију модела машинског учења у различитим контекстима.
- **Хетерогени и шумни подаци:** Употреба машинског учења је најпогоднија за велике и добро куриране скупове података (curated data sets), док подаци о ретким болестима могу бити хетерогени, непотпуни или погрешно означени (нпр. погрешне дијагнозе или погрешно обележавање могу бити уобичајени у студијама ретких болести због ограниченог или еволуирајућег разумевања стања)¹⁶³158. Подаци везани за ретке болести могу потицати из различитих извора, укључујући клиничке картоне, генетско тестирање и анкете пацијената. Ови извори података могу имати различите формате, стандарде квалитета и нивое детаља. Интеграција и хармонизација таквих хетерогених података може бити значајан изазов.
- **Ограничења трошкова и ресурса:** Прикупљање и аотирање података о ретким болестима може бити скупо и дуготрајно. Многе здравствене организације, укључујући и оне у окружењима са ограниченим ресурсима, можда немају приступ ресурсима за улагање, изградњу и одржавање великих скупова података за ретке болести¹⁶⁴159.
- **Осетљиви подаци:** Медицински подаци су веома осетљиви и постоје строга разматрања управљања подацима, њиховог управљања и заштите. Анонимизација и деидентификација података је уобичајена пракса, међутим, мали узорци код ретких болести повећавају вероватноћу идентификације људи путем триангулације¹⁶⁵160. Дељење података о ретким болестима уз обезбеђивање приватности пацијената и поштовање прописа може бити сложено. Потребни су

¹⁶⁰ Nag S, *et al.* 2022 Deep learning tools for advancing drug discovery and development. *3 Biotech*. **12**: 110

¹⁶¹ Steve Nouri. Generative AI Drugs Are Coming. See:

<https://www.forbes.com/sites/forbestechcouncil/2023/09/05/generative-ai-drugs-are-coming/> (accessed September 30 2023)

¹⁶² Banerjee J *et al.* 2023 Machine learning in rare disease. *Nat Methods* 20, 803–814.

(<https://doi.org/10.1038/s41592-023-01886-z>)

¹⁶³ *Ibid.*

¹⁶⁴ The Royal Society interviews with scientists and researchers. 2022 – 2023

¹⁶⁵ The Royal Society interviews with scientists and researchers. 2022 – 2023

додатни ресурси за успостављање поузданих и безбедних истраживачких окружења која омогућавају дељење осетљивих података.

СТРАТЕГИЈЕ ЗА МАКСИМИЗИРАЊЕ УЛОГЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ У ИСТРАЖИВАЊУ РЕТКИХ БОЛЕСТИ

Заједнички напори у дељењу података, стандардизацији и управљању подацима могу отворити пут вештачкој интелигенцији да оствари значајан утицај на проучавање ретких болести и побољшање живота оних који су погођени овим стањима.

- **Независни и интероперабилни регистри пацијената:** Потребна је међуинституционална, међународна сарадња како би се створили велики, централизовани скупови података погодни за истраживања заснована на машинском учењу¹⁶⁶161. Иницијативе за обједињавање података су такође пожељне, као што су регионални или глобални регистри пацијената¹⁶⁷162 који могу проширити приступ релевантним подацима и промовисати стандардизацију и интероперабилност регистара између институција. На пример, Европска платформа за ретке болести објавила је „Скуп заједничких елемената података за регистрацију ретких болести“. Успостављање федеративних инфраструктура за учење, као што је Gaia-X, такође може олакшати дељење осетљивих података¹⁶⁸163.

- **Користите вештачку интелигенцију за проширивање података:** Генеративне технике могу бити ефикасан начин за решавање проблема оскудице података, шума или непотпуности¹⁶⁹164. На пример, стратегије проширивања података или употреба синтетичких података могу се користити за попуњавање непотпуних скупова података вештачким узорцима који повећавају разноликост и репрезентативност скупова података (нпр. решавају изузетке и пристрасности), минимизирајући потребу за личним подацима. Слично томе, приступи рачунарског вида могу се користити за побољшање квалитета и верности података снимања. Нерешени изазови укључују обезбеђивање поузданости и адекватне обуке генеративних модела.

- **Успостављање мрежа за сарадњу са више заинтересованих страна:** Истраживачи ретких болести нагласили су важност сарадње како би се проширио приступ ресурсима и ускладили интереси више заинтересованих страна. Мрежа за ретке болести Азијско-пацифичке економске сарадње је релевантан модел који окупља креаторе политике, академску заједницу и индустрију ради управљања и хармонизације пракси података¹⁷⁰165

¹⁶⁶ Boycott KM *et al.* 2017 International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.* 100, 695–705. (<https://doi.org/10.1016/j.ajhg.2017.04.003>)

¹⁶⁷ Bellgard MI, Snelling T, McGree JM. 2019 RD-RAP: beyond rare disease patient registries, devising a comprehensive data and analytic framework. *Orphanet J Rare Dis* 14, 176. (<https://doi.org/10.1186/s13023-019-1139-9>)

¹⁶⁸ Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. 2021 Opportunities and challenges for machine learning in rare diseases. *Frontiers in Medicine*, 8, 747612 (<https://doi.org/10.3389/fmed.2021.747612>)

¹⁶⁹ Kokosi T, Harron K. 2022. Synthetic data in medical research. *BMJ medicine*, 1. (<https://doi.org/10.1136/bmjmed-2022-000167>)

¹⁷⁰ Global Rare Disease Policy Network. See: <https://www.rarediseasepolicy.org/> (accessed 21 March 2024)

ПОГЛАВЉЕ 2

Интегритет и поузданост истраживања

„Тешко је замислити веће улоге од ових за свет науке. Будуће постојање и друштвена улога [науке] изгледа да зависе од способности истраживача и научних институција да одговоре на кризу, чиме се спречава потпуни губитак поверења у научну стручност од стране цивилног друштва. “

Professor Sabina Leonelli¹⁷¹166

¹⁷¹ Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: “Curiosity, Imagination, and Surprise” of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (<https://doi.org/10.1108/S0743-41542018000036B009>)

Интегритет и поузданост истраживања

Поверење у вештачку интелигенцију је неопходно за њену одговорну употребу у научним истраживањима, посебно зато што научници постају све више зависни од ових технологија¹⁷²167. Ово ослањање се заснива на претпоставци да системи засновани на вештачкој интелигенцији – као и њихове анализе и резултати – могу произвести поуздане и веродостојне налазе, са мало грешака. Међутим, усвајање вештачке интелигенције у научним истраживањима повезано је са изазовима за ригорозност и научни интегритет. Кључна питања укључују недостатак разумевања о томе како модели вештачке интелигенције функционишу, недовољну документацију експеримената и научнике којима недостаје потребна техничка стручност за изградњу модела, њихово тестирање и проналажење грешака. Све већи број нерепродуцибилних студија, које користе технике машинског учења, такође изазива забринутост у вези са изазовима репродукције експеримената заснованих на вештачкој интелигенцији и поузданошћу резултата и открића заснованих на њој¹⁷³168. Ова питања представљају ризик не само за науку, већ и за друштво ако примена непоузданих или неисплативих технологија вештачке интелигенције доведе до штетних исхода¹⁷⁴169

На основу интервјуа и округлог стола о репродуктивности спроведених за овај извештај, уочена су следећа запажања која обухватају јединствене изазове које вештачка интелигенција уноси у вези интегритета и поузданост истраживања.

Изазови репродуктивности у истраживањима заснованим на вештачкој интелигенцији

Репродуктивност се односи на способност независних истраживача да пажљиво испитају резултате истраживачке студије, реплицирају их и репродукују експеримент у будућим студијама¹⁷⁵170. Ако истраживачи развију претерано ослањање на вештачку интелигенцију за анализу података, а притом остану неспособни да објасне како су дошли до закључака и како да репродукују студију¹⁷⁶171, она неће испунити прагове за испитивање и верификацију. Слично томе, ако се резултати не могу верификовати, они могу допринети надуваним очекивањима, преувеличаним тврдњама о тачности или истраживачким исходима заснованим на лажним корелацијама¹⁷⁷172. У случају истраживања заснованог на вештачкој интелигенцији, могућност репродукције студије не подразумева само репликацију методе, већ и могућност репродукције кода, података и услова околине под којима је експеримент спроведен (нпр. рачунарство, хардвер,

¹⁷² Echterhölter A, Schröter J, Sudmann A. 2021 How is Artificial Intelligence Changing Science? Research in the Era of Learning Algorithms. OSF Preprint (<https://doi.org/10.33767/osf.io/28pnx>)

¹⁷³ Sohn E. 2023 The reproducibility issues that haunt health-care AI. Nature. 9 January 2023. See <https://www.nature.com/articles/d41586-023-00023-2> (accessed 21 December 2023)

¹⁷⁴ Sambasivan N *et al.* 2021 “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. (<https://doi.org/10.1145/3411764.3445518>)

¹⁷⁵ Haibe-Kains B *et al.* 2020 Transparency and reproducibility in artificial intelligence. Nature. 586, E14–E16. (<https://doi.org/10.1038/s41586-020-2766-y>)

¹⁷⁶ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/> (accessed 7 May 2024)

¹⁷⁷ Echterhölter A, Schröter J, Sudmann A. 2021 How is Artificial Intelligence Changing Science? Research in the Era of Learning Algorithms. OSF Preprint (<https://doi.org/10.33767/osf.io/28pnx>)

софтвер)^{178,179}173,174 Неуспеси у репродуктивности не само да угрожавају валидност појединачне студије¹⁸⁰175, већ могу утицати и на истраживања спроведена за друге студије, укључујући и оне у другим дисциплинама. На пример, студија коју је водио Центар за статистику и машинско учење на Универзитету Принстон показала је како „цурење података“ у једној студији (водећи узрок грешака у апликацијама машинског учења због грешака у подацима за обуку или карактеристикама модела) може утицати на 294 рада у 17 научних области, укључујући области са високим улогом попут медицине¹⁸¹176. Штавише, ове врсте проблема вероватно неће бити довољно пријављене због фактора као што су необјављени подаци; недовољна документација; одсуство механизма за пријављивање неуспелих експеримената; и велика варијабилност у различитим експерименталним или истраживачким контекстима¹⁸²177.

Непрозирност и природа црне кутије машинског учења

У сржи изазова репродуктивности су непрозирни модели засновани на машинском учењу које не може сваки научник да објасни, протумачи или разуме. Модели машинског учења се обично називају „моделима црне кутије“ (модел који могу да произведу корисне информације и резултате, чак и када истраживачи не разумеју тачно како систем функционише). Непрозрачна природа модела ограничава објашњивост и способност научника да тумаче како модели машинског учења долазе до одређених резултата или закључака¹⁸³178

Објашњива вештачка интелигенција (видети Оквир 1) може помоћи истраживачима да идентификују грешке у подацима, моделима или претпоставкама – ублажавајући изазове као што је пристрасност података – и осигуравајући да ови системи производе висококвалитетне резултате који се могу користити за имплементацију у стварном свету¹⁸⁴179. Ово може постати значајан изазов за научнике који интегришу веома варијабилне и сложене моделе у свој рад, као што су модели дубоког учења, за које се зна да надмашују мање сложене и линеарније и транспарентније моделе. Непрозрачност се повећава када се модели развијају у комерцијалном окружењу. На пример, већину водећих мастер програма учења развијају велике технолошке компаније попут Google, Microsoft, DeepMind. Ови модели су власнички системи и као такви откривају ограничене информације о архитектури свог модела, подацима за обуку и процесима доношења одлука који би побољшали разумевање¹⁸⁵180.

¹⁷⁸ Gundersen O, Gil Y and Aha D. 2018 On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine*. **39**: 56-68. (doi.org/10.1609/aimag.v39i3.2816)

¹⁷⁹ Gunderson O, Coakley K, Kirkpatrick C, and Gil Y. 2022 Sources of irreproducibility in machine learning: A review. *arXiv preprint*. (doi.org/10.48550/arXiv.2204.07610)

¹⁸⁰ McDermott M, Wang S, Marinsek N, Ranganath R, Foschini L, and Ghassemi M. 2021 Reproducibility in machine learning for health research: Still a way to go. *Sci. Transl. Med.* **13**, eabb1655. (doi.org/10.1126/scitranslmed.abb1655)

¹⁸¹ Kapoor S and Narayanan A. 2023 Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. **4**(9) (doi.org/10.1016/j.patter.2023.100804)

¹⁸² Gundersen O, Gil Y and Aha D. 2018 On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine*. **39**: 56-68. (doi.org/10.1609/aimag.v39i3.2816)

¹⁸³ Royal Society. Royal Society response on Reproducibility and Research Integrity. See: <https://royalsociety.org/newsresources/publications/2021/research-reproducibility/> (accessed 7 March 2024)

¹⁸⁴ The Royal Society. 2019 Explainable AI: the basics. See <https://royalsociety.org/topics-policy/projects/explainable-ai/> (accessed 21 December 2023).

¹⁸⁵ Bommasani *et al.* 2021. On the opportunities and risks of foundation models. See: <https://crfm.stanford.edu/assets/report.pdf> (accessed March 21 2024)

„Можда постоји несразмеран проблем са машинским учењем. Дошли смо веома далеко са могућношћу да обрађујемо огромне количине података, користећи софтвер који је веома компетентан и добро развијен. Али мислим да можда многи људи који га користе заправо не разумеју шта раде на начин који можда не важи толико за друге области. То је сложени проблем, где постоји много, много ствари које можете погрешити. Питам се колико људи заиста разуме софтвер који користе.“ **Royal Society roundtable participant**

BOX 1

Објашњивост и интерпретабилност

Објашњивост и интерпретабилност односе се на информације које омогућавају корисницима да разумеју како систем вештачке интелигенције функционише и образложење које стоји иза његових резултата¹⁸⁶181. На пример, у машинском учењу, методе интерпретабилности могу понудити информације о томе „како модел функционише“, док објашњивост одговара на питање зашто се долази до одређених закључака или „шта ми још овај модел може рећи?“¹⁸⁷182. Као што је наведено у извештају Краљевског друштва из 2019. године, „Објашњивост вештачке интелигенције: Основе“¹⁸⁸183, обезбеђивање објашњивости и интерпретабилности у науци може имати следеће предности за поузданост:

- Помаже истраживачима да боље разумеју увиде и обрасце до којих долазе коришћењем сложених модела машинског учења и великих база података.
- Повећава потенцијал научника да извуку увиде из система вештачке интелигенције како би открили потенцијална нова научна открића или продоре¹⁸⁹184.
- Побољшава репродуктивност омогућавајући трећим странама да испитају модел, као и да идентификују и исправе грешке.
- Побољшава преносивост и процену да ли модели могу бити погодни у различитим дисциплинама или контекстима.
- Побољшава одговорност и осигурава да научници могу да понуде оправдање за употребу модела машинског учења¹⁹⁰185.
- У случају научно заснованих апликација које утичу на јавност – од здравља до јавне политике – објашњивост може осигурати да креатори политике и регулатори могу да обезбеде надзор и спрече штету узроковану погрешним предвиђањима или моделима¹⁹¹186

¹⁸⁶ Marcinkevičs, R., Vogt, J. E. 2023. Interpretable and explainable machine learning: A methods-centric overview with concrete examples (<https://doi.org/10.1002/widm.1493>)

¹⁸⁷ Lipton, J. 2018. The mythos of model interpretability. Queue, 16(3), 31–57. (<https://doi.org/10.1145/3236386.3241340>)

¹⁸⁸ The Royal Society. 2019 Explainable AI: the basics. See <https://royalsociety.org/topics-policy/projects/explainable-ai/> (accessed 21 December 2023).

¹⁸⁹ Li Z, Ji J, and Zhang Y. 2023 From Kepler to Newton: Explainable AI for science. *arXiv preprint*. (doi.org/10.48550/arXiv.2111.12210)

¹⁹⁰ McDermid J, Jia Y, Porter Z and Habli I. 2021 Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A*. 379(2207), 20200363. (doi.org/10.1098/rsta.2020.0363)

¹⁹¹ McGough M. 2018 How bad is Sacramento’s air, exactly? Google results appear at odds with reality, some say. *Sacramento Bee*. 7 August 2018. See <https://www.sacbee.com/news/california/fires/article216227775.html> (accessed 21 December 2023).

Компромис између објашњивости и тачности

Једно од главних ограничења објашњивости је оно што се назива компромисом између објашњивости и тачности¹⁹²187. Сматра се да се највећа тачност моделирања (у смислу предвиђања или класификације) за велике модерне скупове података често постиже непрозирним сложеним моделима. Ово је повезано са општим прихватањем непрозирности међу корисницима вештачке интелигенције који користе моделе машинског учења. Конкурентно и брзо усвајање вештачке интелигенције у научним истраживањима још више учвршћује ово прихватање. Тренутни екосистем вештачке интелигенције награђује моделе високих перформанси и конкурентне моделе који су „корисни“ и тачни, а нетранспарентни, приступачни или „прилагођени корисницима“¹⁹³188. Дакле, поставља се питање да ли је непрозирност нормализована као нови статус кво и да ли је објашњивост остварив циљ вредан тежње^{194,195}189,190. Штавише, иако транспарентност може побољшати разумевање, пружање сложених техничких информација о систему не мора увек побољшати способност крајњих корисника да интерагују са системима и разумеју их¹⁹⁶191. Као алтернативу ограничавању истраживања само на објашњиве моделе, неки су предложили већи фокус на побољшању интерпретабилности модела¹⁹⁷192.

Обећавајући приступи за побољшање објашњивости или интерпретабилности укључују:

- Методе објашњиве вештачке интелигенције (XAI) специфичне за различите научне дисциплине: Како корисници из различитих дисциплина интегришу вештачку интелигенцију у свој рад, методе објашњивости постају специфичне за одговарајућу дисциплину и примену. XAI методе се појављују у областима као што су наука о материјалима¹⁹⁸193, биомедицина¹⁹⁹194, наука о Земљи(геонаука)²⁰⁰195, и истраживање животне средине²⁰¹196 за вишеструке сврхе. То укључује унапређење научног разумевања изведеног из вештачке интелигенције (нпр. боље разумевање

¹⁹² Lundberg S and Lee S. 2017 A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4768–4777.

(dl.acm.org/doi/10.5555/3295222.3295230)

¹⁹³ Cartwright H. 2023 Interpretability: Should – and can – we understand the reasoning of machine-learning systems? In: OECD (ed.) *Artificial Intelligence in Science*. OECD. (doi.org/10.1787/a8d820bd-en)

¹⁹⁴ Royal Society roundtable on reproducibility, April 2023.

¹⁹⁵ Birhane A et al. 2023 Science in the age of large language models. *Nat Rev Phys* 5, 277–280.

(doi.org/10.1038/s42254-023-00581-4)

¹⁹⁶ Bell A, Solano-Kamaiko I, Nov O, and Stoyanovich J. 2022 It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. *In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery. 248–266.

(doi.org/10.1145/3531146.3533090)

¹⁹⁷ Miller K. 2021 Should AI models be explainable? That depends. *Stanford University Human-Centered Artificial Intelligence*. See: <https://hai.stanford.edu/news/should-ai-models-be-explainable-depends> (accessed 21 December 2023)

¹⁹⁸ Zhong X et al. 2022 Explainable machine learning in materials science. *NPJ Comput Mater* 8, 204. (doi.org/10.1038/s41524-022-00884-7)

¹⁹⁹ Combi C et al. 2022 A manifesto on explainability for artificial intelligence in medicine. *Artificial intelligence in medicine*. 133, 102423. (doi.org/10.1016/j.artmed.2022.102423)

²⁰⁰ Hanson, B. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. See <https://www.nature.com/articles/d41586-023-03316-8> (accessed 5 March 2024)

²⁰¹ Arashpour M. 2023 AI explainability framework for environmental management research. *Journal of environmental management*. 342, 118149. (doi.org/10.1016/j.jenvman.2023.118149)

физичких принципа и генерисање нових хипотеза²⁰²197); побољшање надзора и спровођења прописа о заштити животне средине; и минимизирање еколошког отиска система вештачке интелигенције²⁰³198.

- **Архитектуре стаклене кутије:** Архитектуре модела стаклене кутије имају за циљ да учине интерне репрезентације података LLM-а транспарентнијим укључивањем механизма пажње, модуларних структура и алата за визуелизацију који могу помоћи у откривању начина на који информације теку кроз слојеве неуронске мреже. Поред тога, проширене технике обуке попут супарничког учења(Adversarial machine learning) и контрастивних примера(contrastive learning) могу испитати границе одлучивања модела. Анализирање када LLM успева или не успева на овим посебним узорцима обуке пружа увид у његов процес расуђивања^{204,205}199,200

- **Графови знања:** Графови знања(Knowledge graphs) су напредна структура података која представља информације у мрежи међусобно повезаних ентитета. Они смањују ослањање на непрозирне статистичке обрасце у подацима за обуку за мастер студије коришћењем великог језичког модела (LLM). Медицинске мастер студије (LLM), на пример, могу искористити онтолошке биомедицинске податке у графовима знања за транспарентно структурирано резонување о болестима и лечењу. Током закључивања, на мастер студијама се консултују графове знања за релевантне чињенице, пружајући утемељен оквир поред суштинског препознавања образаца. Заједничка обука са графовима знања побољшава чињенично резонување мастер студија у овој области и помаже у идентификовању празнина или погрешних схватања кроз ревизије²⁰⁶201.

Баријере које ограничавају репродуктивност

Поред техничких изазова, постоји низ институционалних и друштвених ограничења која спречавају истраживаче да усвоје ригорозније и транспарентније процесе. Табела 1 наводи кључне препреке репродуктивности у истраживањима заснованим на вештачкој интелигенцији²⁰⁷202.

„Једна од ствари која је истинита код моделирања јесте да можете добити готово сваки резултат који желите на основу претпоставки које користите за њихово покретање. Мислим да је ово опасно подручје у којем се наша област креће. Превише се ослањамо на резултате модела и лепе слике које из њих произилазе као репродукција истине.“

Учесник округлог стола Краљевског друштва

²⁰² Zhong X et al. 2022 Explainable machine learning in materials science. *npj Comput Mater* **8**, 204. (doi.org/10.1038/s41524-022-00884-7)

²⁰³ Arashpour M. 2023 AI explainability framework for environmental management research. *Journal of environmental management*. 342, 118149. (doi.org/10.1016/j.jenvman.2023.118149)

²⁰⁴ Lengerich, B J et al. 2023. LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs. arXiv preprint (https://doi.org/10.48550/arXiv.2308.01157)

²⁰⁵ Garrett BL, Rudin C 2023. The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. Cornell Law Review, Forthcoming, Duke Law School Public Law & Legal Theory Series.

²⁰⁶ Gaur, M, Faldu, K, & Sheth, A 2021. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing*, **25**, 51-59.

²⁰⁷ Royal Society roundtable on reproducibility, April 2023.

ТАБЕЛА 1

Баријере за репродуктивност	Примери
Заблуде и претпоставке о ML ²⁰⁸ 203	- Основна претпоставка да су модели машинског учења (ML) инхерентно репродуктивни због свог ослањања на рачунање. - Прекомерно ослањање на излазе засноване на ML и сумњива употреба статистичких техника за ублажавање пристрасности или искључивање неугодних или непожељних резултата.
Рачунарски или услови околине	- Различита хардверска и софтверска окружења могу дати различите резултате. - Репродуктивност у великим размерама подразумева приступ рачунарским капацитетима који омогућавају истраживачима да валидирају сложене моделе машинског учења²⁰⁹204. - Компаније приватног сектора имају боље ресурсе од академске заједнице и могу себи приуштити обуку и валидацију већих модела (нпр. OpenAI-јев GPT-4), док истраживачи у другим секторима не могу²¹⁰205.
Документација и праксе транспарентности	- Недовољна или непотпуна документација о методама истраживања, коду, подацима или рачунарским окружењима. - Растући развој и усвајање мање транспарентних, власничких модела. - Недостатак документације специфичне за дисциплину која се бави препрекама са којима се сусрећу различите области, примене и истраживачки контексти (нпр. документација специфична за здравствену заштиту која се бави смерницама за репродуктивност у истраживању лечења и дијагнозе болести). - Недовољни напори да се документација учини доступном научницима из различитих порекла и са различитим нивоима техничке стручности.
Вештине, обука и капацитети	- Недостатак јасноће у вези са тим ко је одговоран за различите фазе тока рада и мало ресурса за укључивање рада на репродуктивности. - Недостатак обуке за нове кориснике машинског учења и недовољне смернице о ограничењима различитих модела и прикладности различитих техника за примене специфичне за област. - Недостатак алата за стручњаке који нису стручњаци за машинско учење да прате смернице за репродуктивност и идентификују ограничења модела. - Недостатак механизма који олакшавају интердисциплинару сарадњу између научника који немају техничко искуство у вештачкој интелигенцији и рачунарских или научника за податке који имају стручност за унос података, идентификовање грешака и валидацију експеримената.
Подстицаји и култура истраживања	- Мало могућности за напредовање у каријери у академским круговима за улоге потребне за унапређење отворених и репродуктивних истраживања (нпр. Курирање (организација интеграција података из различитих извора) и обрада података; управљање истраживачким подацима; управљање подацима; менаџери истраживања). - Нема подстицаја за објављивање грешака у истраживањима заснованим на машинском учењу (неуспешни резултати) или исправљања - Уски поглед на то који резултати вреде објављивања (нпр. подаци, модели) и ограничене награде за спровођење пракси отворене науке и објављивање извештаја о репродуктивности. - Нема специфичних подстицаја за подстицање употребе и развоја модела које људи могу да интерпретирају када је то могуће²¹¹206.

²⁰⁸ Benjamin D *et al.* 2018 Redefine statistical significance. *Nat Hum Behav* 2, 6–10. (doi.org/10.1038/s41562-017-0189-z)

²⁰⁹ Bommasani *et al.* 2021. On the opportunities and risks of foundation models. See: <https://crfm.stanford.edu/assets/report.pdf> (accessed March 21 2024)

²¹⁰ *Ibid*

²¹¹ Rudin C. 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1,206–215). (doi.org/10.1038/s42256-019-0048-x)

Култура истраживања и репродуктивност

Култура „објави или нестани“ истакнута је као кључни ограничавајући фактор за научнике, у мери у којој награђује број и врсту публикација као неопходне за напредовање у каријери, али не препознаје скупове података, документацију или репродуковане студије као резултате вредне награђивања. Уз неефикасне праксе контроле квалитета и самокорекције унутар часописа, сматра се да тренутни систем објављивања игра значајну улогу у смањењу подстицаја за обављање дуготрајног и колаборативног рада потребног за демонстрацију репродуктивности²¹²207.

Ови изазови су даље истражени у извештају Краљевског друштва из 2018. године, „Култура истраживања: Уградња инклузивне изврности“. Извештај истиче да један од главних подстицаја за ширење резултата истраживања треба да буде корист заједнице у целини и унапређење истраживачког подухвата²¹³208. Такође се разматра вредност транспарентности за уградњу културе интегритета и као средство за заштиту од непотребног дуплирања истраживања.

Репродуктивност у различитим контекстима

Универзални приступ постизању репродуктивности није пожељан. Ово обично захтева висок ниво контроле над условима животне средине и друштвеним условима студије, као и директну репликацију улаза, резултата и метода које могу бити неизводљиве у различитим истраживачким окружењима, културама и контекстима.

Стандардизовани приступ репродуктивности такође може обесхрабрити истраживаче да приступе документацији и извештавању са рефлексивног становишта које се бави варијабилношћу и идиосинкратичнијим аспектима научног истраживања²¹⁴209. Модели који се не могу генерализовати у различитим контекстима могу, на пример, понудити вредне увиде у вези са извором варијација и зашто су оне важне. У контексту здравствене заштите, локални услови (нпр. протоколи пријема, лабораторијско тестирање, управљање евиденцијом или интеракције између лекара и пацијента) болнице могу значајно променити резултате.

Алтернатива стандардизованом приступу побољшању репродуктивности је контекстуалнији приступ документацији и истраживачким протоколима који обухвата варијабилност и пружа увид у локалну адаптацију модела у различитим контекстима²¹⁵210. Овај приступ има потенцијал да подржи истраживаче који желе да прилагоде моделе, уместо да извозе или увозе моделе који се не преносе добро у различите географске и културне „контексте открића“²¹⁶211

²¹² Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: “Curiosity, Imagination, and Surprise” of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (<https://doi.org/10.1108/S0743-41542018000036B009>)

²¹³ The Royal Society. 2018 Research culture: embedding inclusive excellence. See <https://royalsociety.org/topicspolicy/publications/2018/research-culture-embedding-inclusive-excellence/> (accessed 21 December 2023)

²¹⁴ Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: “Curiosity, Imagination, and Surprise” of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (<https://doi.org/10.1108/S0743-41542018000036B009>)

²¹⁵ Miller K. 2022 Healthcare algorithms don’t always need to be generalizable. *Stanford University Human-Centered Artificial Intelligence*. See <https://hai.stanford.edu/news/healthcare-algorithms-dont-always-need-be-generalizable> (accessed 21 December 2023).

²¹⁶ Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: “Curiosity, Imagination, and Surprise” of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. **36B**, 129-146 (<https://doi.org/10.1108/S0743-41542018000036B009>)

„Криза [репродуктивности] или изазов машинског учења односи се на то ко је укључен у одређивање како се репродуктивност дефинише и за кога ће ова дефиниција бити корисна? Да ли примењујемо контекстуално знање и ситуационо разумевање за шта ће се технологија користити? У којој мери ће подржати различите истраживачке заједнице да побољшају поузданост својих експеримената?“

Учесник округлог стола Краљевског друштва

BOX 2

Робусност и генерализација у машинском учењу

Робусност модела се односи на способност модела да тачно функционише у различитим контекстима²¹⁷212. Док су модерни модели машинског учења оптимизовани за обављање уских и специфичних задатака, развој вештачке интелигенције укључује развој капацитета за „генерализацију“ или пренос учења из задатка обуке у нове примене. Без генерализације, модели вештачке интелигенције могу добро функционисати на подацима за обуку, али не и на новим, непознатим подацима или у новим истраживачким окружењима. Репродуктивност игра главну улогу у осигуравању да стране из различитих истраживачких окружења могу реплицирати експеримент и тестирати генерализацију

Унапређење транспарентности и поузданости

Изазови у вези са поузданошћу навели су научнике да развију истраживачке протоколе, стандарде, алате и праксе отворене науке како би се осигурала транспарентност и научна ригорозност у истраживањима заснованим на вештачкој интелигенцији²¹⁸213. То укључује:

Подстицање објављивања извештаја о репродуктивности

- Претходна регистрација и регистровани извештаји. Иницијативе за објављивање методологија, које промовише Центар за отворену науку, могу побољшати транспарентност истраживачких студија, подстичући истраживаче да документују свој план истраживања пре него што спроведу даља истраживања и поднесу методологију на рецензију²¹⁹214.
- Сервери за претпринт. Повећана употреба сервера за претпринт (нпр. као што је bioRxiv од стране биолошких и биомедицинских заједница) може играти улогу у олакшавању комуникације о успешним и неуспешним резултатима репликације.
- Велики изазови. На пример, изазов репродуктивности машинског учења позива учеснике да репродукују радове објављене на једанаест водећих конференција о машинском учењу и објаве извештај о репродуктивности вођен од стране заједнице који документује налазе²²⁰215.

²¹⁷The Royal Society. 2017 Machine Learning: The power and promise of computers that learn by example. See <https://royalsociety.org/topics-policy/projects/machine-learning/> (accessed 21 December 2023)

²¹⁸ Birhane A et al. 2023 Science in the age of large language models. *Nat Rev Phys* 5, 277–280. (doi.org/10.1038/s42254-023-00581-4)

²¹⁹ Center for Open Science. What is preregistration? See <https://www.cos.io/initiatives/prereg> (accessed 21 December 2023).

²²⁰ Papers With Code. ML Reproducibility Challenge 2022. See <https://paperswithcode.com/rc2022> (accessed 21 December 2023).

Упутство за израду документације и праћење пракси отворене науке

- Контролне листе и протоколи за репродуктивност. Примери укључују Контролну листу за репродуктивност машинског учења²²¹216, Контролну листу за вештачку интелигенцију у медицинском снимању (CLAIM)²²²217 или контролну листу REFORMS независну од терена²²³218, коју су развили стручњаци из рачунарства, математике, друштвених наука и здравствених истраживања. Ово олакшава усклађеност и документовање вишеструких димензија репродуктивности.
- Стандарди заједнице за документацију. Развој стандарда заједнице специфичних за домен, као што је TRIPOD-AI²²⁴219, пружа смернице о томе како документовати, извештавати и репродуковати студије модела предвиђања засноване на машинском учењу у здравственим истраживањима. Заједница синтетичке биологије и геномике је такође дефинисала експерименталне протоколарне стандарде и документацију геномског тока рада да би се побољшала репродуктивност^{225,226}220,221
- Објављивање листова података и картица модела. Индустрија може играти важну улогу у објављивању информација које пружају увид у то шта модел ради; његову циљну публику; намењену употребу; потенцијална ограничења; метрике поузданости; и информације о архитектури модела и подацима за обуку. Meta²²⁷222, Google²²⁸223 и Hugging Face²²⁹224 су објавили различите итерације картица модела.
- Документација свесна контекста. Укључивање различитих актера у дефинисање начина дефинисања репродуктивности; промовисање механизма извештавања који експлицитно адресирају контекстуалне улазе и изворе варијација; и документовање како локална или тимска култура утиче на имплементацију²³⁰225

Алати и платформе за сарадњу и приступ

- Онлајн платформе за сарадњу и репозиторијуми који олакшавају дељење скупова података; верзија софтвера; алгоритама; токова рада; и метода. Примери укључују CodaLab²³¹226 и OpenML²³²227.

²²¹ McGill School of Computer Science. The Machine Learning Reproducibility Checklist v2.0.

See: <https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf> (accessed 21 December 2023).

²²² Mongan J, Moy L, and Kahn C. 2020 Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology. Artificial intelligence*, 2(2), e200029. (doi.org/10.1148/ryai.2020200029)

²²³ Reporting standards for ML-based science. See: <https://reforms.cs.princeton.edu/> (accessed 21 December 2023).

²²⁴ Collins G et al. 2021 Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ open*, 11(7), e048008. (doi.org/10.1136/bmjopen-2020-048008)

²²⁵ Lin X. 2020 Learning Lessons on Reproducibility and Replicability in Large Scale Genome-Wide Association Studies. *Harvard Data Science Review*. 2. (doi.org/10.1162/99608f92.33703976)

²²⁶ Kanwal S et al. 2017 Investigating reproducibility and tracking provenance – A genomic workflow case study. *BMC Bioinformatics* 18, 337. (doi.org/10.1186/s12859-017-1747-0)

²²⁷ Meta. 2022 System Cards, a new resource for understanding how AI systems work. *Meta*. 23 February 2022. See <https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/> (accessed 21 December 2023).

²²⁸ Google. Model Cards. See <https://modelcards.withgoogle.com/about> (accessed 21 December 2023).

²²⁹ Hugging Face. Model Cards. See <https://huggingface.co/docs/hub/model-cards> (accessed 21 December 2023).

²³⁰ Leonelli S. 2018 Rethinking reproducibility as a criterion for research quality. In a symposium on Mary Morgan: “Curiosity, Imagination, and Surprise” of Research in the History of Economic Thought and Methodology. *Emerald Publishing Limited*. January 2018. 36B, 129-146 (<https://doi.org/10.1108/S0743-41542018000036B009>)

²³¹ CodaLab. CodaLab Worksheets. See <https://worksheets.codalab.org/> (accessed 21 December 2023).

²³² OpenML. See <https://www.openml.org/> (accessed 21 December 2023).

- Интерактивни алати без кода. Развој приступачних алата за вештачку интелигенцију може помоћи у изградњи поверења и укључивању корисничке стручности у дизајн и процену параметара модела. Примери укључују интерактивне контролне табле²³³228 и решења без кода која користе кориснички пријатељске интерфејсе за приказивање скупова података, метрика поузданости и обучавање модела новим примерима. Алати без кода и „гранични модели“ такође се могу користити у регионима без приступа интернету са ограничењима у функционалности²³⁴229

²³³ Morris M. 2023 Scientists’ perspectives on the potential for generative AI in their fields. *Google Research*. (doi.org/10.48550/arXiv.2304.01420)

²³⁴ Korot E *et al.* 2021 Code-free deep learning for multi-modality medical image classification. *Nat Mach Intell.* **3**, 288–298. (doi.org/10.1038/s42256-021-00305-2)

ПОГЛАВЉЕ 3

ИСТРАЖИВАЧКЕ ВЕШТИНЕ И ИНТЕРДИСЦИПЛИНАРНОСТ

“Долазимо до размера података и сложености, где не можете све сами да урадите, или можете, али ћете бити ограничени. Ако желите да будете успешни, морате да разумете како да сарађујете и ко-стварате са људима. А то је тешко. Болно је и спорије је, али потенцијално има већи утицај, зар не? [...] Ако желите брзе резултате, урадите то сами. Ако желите ствари које имају утицај, морате да сарађујете и радите ствари другачије.”

Royal Society roundtable participant

ИСТРАЖИВАЧКЕ ВЕШТИНЕ И ИНТЕРДИСЦИПЛИНАРНОСТ

Успешна примена вештачке интелигенције у научним истраживањима и њено превођење у вредност у стварном свету могу захтевати интердисциплинарне вештине и знања²³⁵230. Интердисциплинарна истраживања (IDR) обухватају активности које интегришу више од једне дисциплине са циљем стварања новог знања или решавања заједничког проблема²³⁶231. Компјутерским научницима је потребна стручност у области да би дизајнирали одговарајуће моделе, док је стручњацима у области потребна стручност у области вештачке интелигенције да би искористили те алате за своја истраживања. Ова интердисциплинарна сарадња може олакшати размену знања и покренути иновативна решења за сложене глобалне проблеме као што су климатске промене, губитак биодиверзитета и епидемије^{237,238}232,233. Ово поглавље се ослања на увиде добијене са округлог стола који је организовало Краљевско друштво о улози интердисциплинарности у обезбеђивању напретка научних истраживања заснованих на вештачкој интелигенцији²³⁹234

Изазови за интердисциплинарност у истраживањима заснованим на вештачкој интелигенцији

Интердисциплинарна сарадња суочава се са различитим препрекама, укључујући следеће:

1. Изоловане академске дисциплине и истраживачке културе

Дисциплине често функционишу унутар различитих „епистемичких култура“ које обухватају норме; методологије; теоријске оквире; евалуацију; и моделе финансирања²⁴⁰235. Успостављање заједничког језика захтева упорне напоре и иницијативу за премошћавање термилошких, парадигматских и когнитивних празнина²⁴¹236. Интердисциплинарни центри могу усвојити интердисциплинарни приступ истраживањима заснованим на вештачкој интелигенцији. На пример, Центар за проучавање егзистенцијалног ризика на Универзитету у Кембриџу проучава биолошке, еколошке, глобално праведне и екстремне технолошке ризике вештачке интелигенције^{242,243}237,238. Слично томе, докторски програми обуке, као што је UKRI Центар за докторску обуку у области еколошке интелигенције²⁴⁴239, могу неговати интердисциплинарну културу пружајући докторантима интердисциплинарну обуку у етици, управљању и одговорним иновацијама вештачке интелигенције

²³⁵ Eguíluz, V M, Mirasso, C R, & Vicente, R 2021. Fundamentals and Applications of AI: An Interdisciplinary Perspective. *Frontiers in Physics*, 8. (<https://doi.org/10.3389/fphy.2020.633494>)

²³⁶ Weber, C T, & Syed, S. 2019. Interdisciplinary optimism? Sentiment analysis of Twitter data. *Royal Society open science*, 6. (<https://doi.org/10.1098/rsos.190473>)

²³⁷ The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023

²³⁸ Wang H *et al* 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620. 47-60. (<https://doi.org/10.1038/s41586-023-06221-2>)

²³⁹ The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023

²⁴⁰ Zeller F, Dwyer L. 2022 Systems of collaboration: challenges and solutions for interdisciplinary research in AI and social robotics. *Discover Artificial Intelligence*, 2.12. (<https://doi.org/10.1007/s44163-022-00027-3>)

²⁴¹ The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, 2023.

²⁴² University of Cambridge. Centre for the study of existential risk. See <https://www.cser.ac.uk/> (accessed 13 December 2023)

²⁴³ The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023

²⁴⁴ University of Exeter. Environmental Intelligence: Data Science & AI for Sustainable Futures. See <https://www.exeter.ac.uk/research/eicdt/> (accessed 22 February 2024)

2. Изоловане инфраструктуре података

Поједине дисциплине прикупљају велике количине података а њихова изолована инфраструктура ограничава знање и дељење података²⁴⁵240. Интегрисани системи података и конзорцијуми могу ублажити вишак података, стандардизовати обраду података и олакшати истраживачима приступ подацима, уз истовремено смањење дуплирања напора. Примери укључују Службу за податке о животној средини, мрежу која пружа податке и алате из области науке о животној средини у Великој Британији за интердисциплинарну анализу²⁴⁶241 и Платформу за податке о океанима, која сакупља глобалне податке о океанима у облаку, превазилазећи ограничења складиштења и рачунарства²⁴⁷242.

3. Различити модели објављивања у научним дисциплинама

Дивергентни модели објављивања ограничавају интердисциплинарну сарадњу у истраживању вештачке интелигенције. Примери укључују разлике у местима објављивања (публикације на конференцијама наспрам публикација у часописима), стиловима објављивања (публикације једног аутора наспрам публикација групе) и нивоима објављивања (отворена наспрам затворене науке), што утиче на подстицаје и спремност за сарадњу у истраживањима заснованим на вештачкој интелигенцији²⁴⁸243

Нови модел објављивања на конференцијама о вештачкој интелигенцији, вођен честим роковима, убрзава обим рада, што доводи до циклуса честих конференција, као што су Конференција о системима за обраду неуронских информација (NeurIPS) и Међународна конференција о машинском учењу (ICML)²⁴⁹244. Ово може довести до великог броја радова којима недостаје дубина и квалитет, што је у супротности са итеративним процесом рецензирања часописа²⁵⁰245. Тајне праксе (нпр. затворени подаци и модели) такође су у супротности са принципима отворене науке. Овај модел се значајно разликује од традиционалних публикација у академским часописима и може ограничити поуздан IDR у истраживањима заснованим на вештачкој интелигенцији.

Хибридни модел часописа и конференције где радови пролазе кроз темељну рецензију у часописима са кратким роком обрта, као што је Часопис за истраживање машинског учења (JMLR), пре презентација на конференцији, могао би да подстакне резултате вишег квалитета и олакша сарадњу између дисциплина које вреднују различите норме објављивања²⁵¹246. Постизање овога вероватно ће захтевати приступ са више заинтересованих страна, укључујући агенције за финансирање, истраживачке институције и заједнице конференција о вештачкој интелигенцији, како би се уравнотежила дисеминација најсавременијих технологија са дубљим прегледом.

4. Сарадња са истраживачима који нису из области науке, технологије, инжењерства и математике (STEM)

Усклађивање квантитативних и квалитативних метода може представљати изазове у току рада за интердисциплинарну сарадњу између истраживача из области науке, технологије, инжењерства и математике (STEM) и истраживача који нису из области(STEM) попут уметности, хуманистичких и друштвених наука. Штавише, истраживачи који нису из области науке, технологије

²⁴⁵National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>.

²⁴⁶ UKRI NERC Environmental Data Service. See: <https://eds.ukri.org/>

²⁴⁷ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>

²⁴⁸The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023

²⁴⁹ Bengio Y. 2020 Time to rethink the publication process in machine learning. See <https://yoshuabengio.org/2020/02/26/time-to-rethink-the-publication-process-in-machine-learning/> (accessed 10 January 2024)

²⁵⁰*Ibid.*

²⁵¹ Slow Science. See <http://slow-science.org/> (accessed 10 January 2024)

и математике суочавају се са додатним препрекама везаним за могућности финансирања и недостатак „правог интердисциплинарног укључивања“, што може ограничити смислену, одрживу сарадњу²⁵²247.

STEAM приступ, где је уметност директно укључена у STEM, може помоћи уметности да допуни STEM истраживање²⁵³248. Интердисциплинарни програми као што су Институт за интердисциплинарну науку о подацима и вештачку интелигенцију на Универзитету у Бирмингему²⁵⁴249 и Колеџ интегративних наука и уметности (CISA) на Државном универзитету Аризоне²⁵⁵250 могу да реше овај јаз подржавајући истраживаче да раде у различитим дисциплинама²⁵⁶251.

Сарадња између стручњака за вештачку интелигенцију и истраживача у областима које нису из области STEM може понудити могућности, укључујући²⁵⁷252:

- Усвајање метода вођених вештачком интелигенцијом у областима које нису из области STEM (нпр. генеративна вештачка интелигенција у уметности и фотографији)²⁵⁸253;
- Искоришћавање креативности и стручности уметника за истраживања заснована на вештачкој интелигенцији (нпр. дизајн корисничког искуства)²⁵⁹254;
- Укључивање перспектива друштвених и хуманистичких наука за одговорну вештачку интелигенцију, безбедност вештачке интелигенције и дискусије о етици истраживања²⁶⁰255.
- Усвајање метода партиципативног истраживања које нису везане за STEM, укључујући грађанску науку и отворене дијалоге, ради повећања поверења, транспарентности и инклузивности у истраживањима заснованим на вештачкој интелигенцији²⁶¹256

„Од студената се очекује да обављају одређене 'службене задатке'. То је важан посао. Ми га називамо 'службеним радом' јер је то заправо служба науци, али није славна, није гламурозно, није довољно призната. И на крају, када тражите академске послове, ове ствари се не цене.“

Учесник интервјуа са Краљевским друштвом

²⁵² The Royal Society interviews with scientists and researchers. 2022 - 2023

²⁵³ The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies.

See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/> (accessed 21 December 2023).

²⁵⁴ University of Birmingham. The Institute for Interdisciplinarity Data Science and AI.

See <https://www.birmingham.ac.uk/research/data-science/index.aspx> (accessed 11 January 2024)

²⁵⁵ Arizona State University. See: <https://news.asu.edu/20230407-university-news-asu-college-integrative-sciencesarts-reorganizes-3-new-schools>. (accessed 13 December 2023)

²⁵⁶ The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies.

See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/> (accessed 21 December 2023).

²⁵⁷ The Royal Society roundtable on the role of interdisciplinarity in AI for scientific research, June 2023.

²⁵⁸ Wu T, Zhang, SH. 2024 Applications and Implication of Generative AI in Non-STEM Disciplines in Higher Education. In: Zhao, F., Miao, D. (eds) AI-generated Content. AIGC 2023. Communications in Computer and Information Science, vol 1946. Springer, Singapore. (doi.org/10.1007/978-981-99-7587-7_29)

²⁵⁹ The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies.

See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/> (accessed 21 December 2023).

²⁶⁰ Zeller F, Dwyer L. 2022 Systems of collaboration: challenges and solutions for interdisciplinary research in AI and social robotics. *Discover Artificial Intelligence*, 2. 12. (<https://doi.org/10.1007/s44163-022-00027-3>)

²⁶¹ UNESCO Recommendation on Open Science. 2021. See: <https://www.unesco.org/en/legal-affairs/recommendationopen-science> (accessed 6 February 2024)

Нове истраживачке вештине у истраживањима заснованим на вештачкој интелигенцији

Повећана примена вештачке интелигенције истиче потребу за основним вештинама вештачке интелигенције и података у свим истраживачким областима²⁶²257. Иако напредна стручност у области података и програмирања можда није увек потребна, нити ће методе вештачке интелигенције бити неопходне за све истраживачке области, свест о најновијим алатима и техникама може повећати доступност и користи за науку. Следећа подручја нових истраживачких вештина извучена су из интервјуа и округлих столова Краљевског друштва у којима су учествовали научници из целе академске заједнице и индустрије.

1. Специјализоване вештине у вези са подацима

Потражња за специјализованим вештинама у вези са подацима је порасла како све више истраживачких области усваја вештачку интелигенцију. Предвиђа се да ће научници и инжењери података имати најбржи раст између 2023. и 2027. године²⁶³258, а улоге које укључују чуваре података и стручњаке за репродуктивност добијају на вредности како се појављују изазови вештачке интелигенције (видети поглавље 2). Међутим, критични задаци управљања подацима, укључујући курирање, чишћење и обезбеђивање квалитета, често су потцењени као „услужни рад“²⁶⁴259. Иако побољшање писмености у вези са подацима у свим дисциплинама може помоћи усвајању вештачке интелигенције, брзи напредак алата вештачке интелигенције ризикује да надмаши обуку о ефикасним праксама у вези са подацима.

2. Обука за вештачку интелигенцију

Јачина вештина може такође укључивати недостатак разумевања изазова везаних за пристрасност, репродуктивност и захтеве за подацима приликом коришћења модела вештачке интелигенције. Усавршавање стручњака у научним областима је неопходно за осмишљавање иновативних апликација вештачке интелигенције. Организације попут Cambridge Spark-а пружају образовне ресурсе за науку о подацима, решавајући јаз вештина кроз стручно усавршавање, корпоративну обуку или кампове за обуку вештина²⁶⁵260. Поред тога, ЕУ подстиче образовање и вештине кроз разне иницијативе у оквиру програма Хоризонт Европа, као што су Европски институт за иновације и заједнице технолошког знања и иновација (EIT KICs), Европски савет за иновације и програм Еразмус+²⁶⁶261

„Као научник који стари и није биоинформатичар, сматрам да су многе од ових ствари прилично непробојне. Често немате самопоуздања да нешто пронађете и биће вам потребне друге колеге да се заиста упусте у то и детаљно погледају.“

Учесник интервјуа Краљевског друштва

²⁶²The Royal Society. 2019. Dynamics of data science skills: How can all sectors benefit from data science talent. See <https://royalsociety.org/-/media/policy/projects/dynamics-of-data-science/dynamics-of-data-science-skills-report.pdf>(accessed 6 January 2024)

²⁶³ World Economic Forum. 2023 The Future of Jobs Report 2023. See https://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf (accessed 30 January 2024)

²⁶⁴ The Royal Society roundtable on reproducibility, April 2023

²⁶⁵ Cambridge Spark. See <https://www.cambridgespark.com/> (accessed 1 August 2023)

²⁶⁶ Petkova, D, Roman, L. 2023 AI in science: Harnessing the power of AI to accelerate discovery and foster innovation— *Policy brief*, Publications Office of the European Commission, Directorate-General for Research and Innovation.(doi/10.2777/401605)

3. Обука о етици вештачке интелигенције

Истраживачи усвајају нове технике вештачке интелигенције, као што је генеративна вештачка интелигенција, без потпуног разумевања етичких импликација. То је због недостатка доказа о потенцијалним ризицима и ограничене обуке о етици вештачке интелигенције. (Види поглавље 5)²⁶⁷262. Сарадња са етичарима и обука о етици вештачке интелигенције могу помоћи у превазилажењу овог јаза. Ово истражују организације као што је Монреалски институт за етику вештачке интелигенције, међународна непрофитна организација, која има за циљ да „демократизује писменост о етици вештачке интелигенције“ пружањем приступачних ресурса, укључујући живи речник и брифинг о етици вештачке интелигенције²⁶⁸263. Укључивање истраживача у активности осигурања вештачке интелигенције такође може допринети изградњи вештина за идентификацију рањивости и ризика у системима вештачке интелигенције. Примери укључују црвено тимовање (red teaming) (видети оквир 3) или обуку о принципима као што је супарничко машинско учење²⁶⁹264

4. Системи и вештине „човек у петљи“

Вештачка интелигенција може побољшати вештине истраживача, помоћи у задацима или аутоматизовати процесе²⁷⁰265 (Види поглавље 1). Такође може играти улогу у подржавању људског расуђивања и креативности у научним подухватима. Дефинисање комплементарних улога за људе и вештачку интелигенцију ради подршке научним истраживањима и преквалификације за аутоматизацију (као што се види у студијама случајева потпуне лабораторијске аутоматизације²⁷¹266) биће неопходно. Ова транзиција такође истиче потребу за системима „људски интеракцијски систем“ за контролу квалитета, осигурање квалитета и прилагођавање променљивој динамици радног процеса²⁷²267.

„Одувек сам користио технологије машинског учења, али нисам размишљао о лошој употреби вештачке интелигенције. Никада у свом животу као научник нисам имао предавање о етици, а то је било у последњих 30 година мог образовног живота. То показује да смо погрешили у последњих 30 година.“

Учесник округлог стола Краљевског друштва

²⁶⁷ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

²⁶⁸ Montreal AI Ethics Institute. See <https://montrealethics.ai/>. (accessed 26 February 2024.)

²⁶⁹ The Royal Society. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation. See <https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation>

²⁷⁰ OECD. 2023. Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research, *OECD Publishing, Paris* (<https://doi.org/10.1787/a8d820bd-en>).

²⁷¹ Archetti C, Montanelli A, Finazzi D, Caimi L, Garrafa E. Clinical laboratory automation: a case study. *J Public Health Res.* 2017;6(1):881. (doi: 10.4081/jphr.2017.881)

²⁷² Al Naam YA et al 2022 The Impact of Total Automaton on the Clinical Laboratory Workforce: A Case Study. *J Health Leadersh.* 9;14:55-62. (doi:10.2147/JHL.S362614)

Увиди из вежбе *црвеног тимирања* Краљевског друштва и Хумане интелигенције о садржају дезинформација генерисаном вештачком интелигенцијом²⁷³268

Црвено тимирање се односи на процес активног идентификовања потенцијалних слабости, начина отказа, пристрасности или других ограничења у моделу, технологији или процесу тако што постоје групе које „нападају“.

У припреми за Глобални самит о безбедности вештачке интелигенције у Великој Британији 2023. године, Краљевско друштво и Хумане интелигенције окупили су 40 постдипломаца из области здравствених и климатских наука како би испитали како потенцијалне рањивости у LLM-овима (Мета-ина Llama 2) могу омогућити производњу научних дезинформација.

Преузимајући различите улоге „актера дезинформација“, учесници су тестирали заштитне ограде модела у вези са темама заразних болести и климатских промена. За мање од два сата, открили су забрињавајуће рањивости, укључујући немогућност модела да пренесе научну несигурност и његово ослањање на сумњиве или фиктивне изворе.

Иако су заштитне ограде спречиле неке уобичајене трендове дезинформација, попут оних везаних за COVID-19, учесници су и даље били у стању да генеришу резултате који су искривљавали проверљиве научне чињенице, доводећи до погрешних закључака.

Вежба је показала вредност укључивања стручњака из домена у процене безбедности вештачке интелигенције пре примене. Њихова научна стручност им је омогућила да тестирају системе на начине који су открили критичне недостатке. Учесници су такође изразили оптимизам у погледу будућности заштитних ограда дезинформација у области мастер студија у области учења (LLM) и веће поверење у коришћење LLM-ова у сопственим истраживањима. Њихови увиди сугеришу да би црвено тимовање могло играти улогу у побољшању писмености у области вештачке интелигенције унутар научне заједнице.

Studija slučaja 2

AI i nauka o materijalima

Наука о материјалима је област у којој технике вештачке интелигенције и машинског учења имају потенцијал да буду трансформативне, са широким друштвеним користима, под условом да постоји одговарајућа инфраструктура подршке.

Потреба за развојем напредних материјала, од нових материјала за батерије за складиштење енергије до катализатора за стварање биоразградиве пластике, била је покретач нових технологија. Историјски гледано, овај сложени процес се у великој мери ослањао на научново претходно

²⁷³ The Royal Society. 2024. Red teaming large language models (LLMs) for resilience to scientific disinformation. See <https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/>

знање, интуицију или случајности захтевајући његову сналажњивост у процењивању неке од 10-ак трилиона могућих хемијских комбинација.

Међутим, вештачка интелигенција и машинско учење сада убрзавају откривање и оптимизацију материјала²⁷⁴269. Ове технике брзо проверавају кандидате, предвиђају структуре и нуде предложене промене које би превазилазиле ручне приступе²⁷⁵270. Заузврат, токови рада засновани на вештачкој интелигенцији могу довести до временске ефикасности, омогућавајући да више идеја напредује од зачећа до комерцијализације у временском периоду година уместо деценија²⁷⁶271

Дизајн и предвиђање материјала

Моделирање и симулација су добро успостављене технике у науци о материјалима. Теорија функционалне густине (Density Functional Theory-DFT)²⁷⁷272 је једна од најчешћих метода моделирања и важан је алат у моделирању материјала. Омогућава прецизна прорачунавања понашања материјала, иако не функционише добро за одређене класе материјала или за одређена својства која су важна за понашање материјала. Такође је рачунски скупа метода и, као таква, ограничена је на материјале ниске до средње сложености²⁷⁸273. Друге технике као што су Монте Карло симулација и молекуларна динамика се такође често користе, али су слично рачунски скупе за сложене материјале²⁷⁹274.

Технике вештачке интелигенције и машинског учења могу се користити за предвиђање структуре и својстава материјала. Пример за то је коришћење генеративних алгоритама и основних модела за предвиђање који материјали би могли показати пожељна својства²⁸⁰275. Међутим, тренутна база знања о материјалима је велика и разнолика, а подаци су често непотпун, недоследно форматирани и лоше означени²⁸¹276. Значајне количине података о материјалима су заробљене у часописима без отвореног приступа или се никада уопште не објављују. Штавише, подаци из негативних или неуспешних експеримената се рутински не објављују.

Последњих година развијено је неколико база података о материјалима са циљем агрегирања података у конзистентним форматима који се затим могу користити за даља истраживања. Примери укључују базе података Materials Project²⁸²277 и Aflow²⁸³278 (које обе

²⁷⁴ Davies *et. al.* 2016 Computational screening of all stoichiometric inorganic materials. *Chem.* **1**, 617-627. (<https://doi.org/10.1016/j.chempr.2016.09.010>).

²⁷⁵ Pyzer-Knapp *et. al.* 2023 Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials.* **8**, 84. (<https://doi.org/10.1038/s41524-022-00765-z>).

²⁷⁶ Materials Genome Initiative. About the Materials Genome Initiative. See <https://www.mgi.gov/about> (accessed 14 July 2023).

²⁷⁷ Argaman N, Makov G. 2000 Density functional theory: An introduction. *American Journal of Physics.* **68**, 69-79. (<https://doi.org/10.1119/1.19375>).

²⁷⁸ Alberi *et. al.* 2019 The 2019 materials by design roadmap. *Journal of Physics D: Applied Physics.* **52**, 013001. (<https://doi.org/10.1088/1361-6463/aad926>).

²⁷⁹ Tao Q, Xu P, Li M, Lu W. 2021. Machine learning for perovskite materials design and discovery. *npj Computational Materials.* **7**, 23. (<https://doi.org/10.1038/s41524-021-00495-8>)

²⁸⁰ Ross *et. al.* 2022 Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence.* **4**, 1256-1264. (<https://doi.org/10.1038/s42256-022-00580-7>).

²⁸¹ Liu Y, Zhao T, Ju W, Shi S. 2017 Materials discovery and design using machine learning. *Journal of Materiomics.* **3**, 159-177. (<https://doi.org/10.1016/j.jmat.2017.08.002>).

²⁸² Jain *et. al.* 2013 The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials.* **1**, 011002. (<https://doi.org/10.1063/1.4812323>)

²⁸³ Curtarolo *et. al.* 2012 AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab-initio* calculations. *Computational Materials Science.* **58**, 227-235. (<https://doi.org/10.1016/j.commatsci.2012.02.002>).

садрже израчуната својства) и базу података Inorganic Crystal Structure Database (ICSD)²⁸⁴279 и базу података High Throughput Experimental Materials (HTEM)²⁸⁵280 (које су обе примери експерименталних база података). Постоје и алати који помажу у креирању и анализи скупова података о материјалима, као што су NOMAD²⁸⁶281, ChemML²⁸⁷282 и atomate²⁸⁸283. Ови скупови података, који могу бити значајне величине (нпр. база података Materials Project тренутно садржи податке за више од 150.000 материјала), олакшали су употребу машинског учења (ML) за откривање материјала. Последњих година било је неколико прича о успеху коришћења машинског учења за откривање материјала, а неки примери су наведени у Табели 2. Различите технике машинског учења и вештачке интелигенције, укључујући генеративну вештачку интелигенцију, коришћене су за идентификацију материјала са жељеним својствима за широк спектар примена. Ове су интегрисане са утврђеним техникама као што су DFT (Density functional theory), прорачуни стабилности и експерименти како би се сузио круг предвиђених материјала²⁸⁹284. Одрживост предложених материјала такође би се могла користити као циљ за предиктивне моделе²⁹⁰285, како би се спречило да нови, сложенији материјали буду тежи за рециклирање или безбедно одлагање.

ТАБЕЛА 2 : Примери машинског учења у откривању материјала

Истраживачи	Резултати
Lyngby et. al. ²⁹¹ 286	Предвиђено је 11.630 нових, стабилних 2Д материјала.
Yao et. al. ²⁹² 287	Пронађене су 2 нове „инварне легуре“ које имају ниско термичко ширење и могу бити корисне за неколико примена
Vasylenko et. al. ²⁹³ 288	Идентификована су 4 нова материјала, укључујући материјале који имају пожељна својства за употребу у чврстим батеријама
Sun et. al. ²⁹⁴ 289	Приступ за претходну проверу нових органских фотонапонских материјала.
Stanev et. al. ²⁹⁵ 290	Идентификовано је >30 потенцијалних високотемпературних суперпроводних материјала.

²⁸⁴ Physical Sciences Data-Science Service. ICSD. See <https://www.psd.ac.uk/icsd> (accessed 14 July 2023).

²⁸⁵ Zakutayev et. al. 2018 An open experimental database for exploring inorganic materials. *Scientific Data*. 5, 180053. (<https://doi.org/10.1038/sdata.2018.53>).

²⁸⁶ Draxl C, Scheffler M. 2019 The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*. 2, 036001. (<https://doi.org/10.1088/2515-7639/ab13bb>).

²⁸⁷ ChemML. See <https://hachmannlab.github.io/chemml/> (accessed 14 July 2023).

²⁸⁸ Atomate. See <https://atomate.org/> (accessed 14 July 2023).

²⁸⁹ DeCost et. al. 2020 Scientific AI in materials science: a path to a sustainable and scalable paradigm. *Machine Learning: Science and Technology*. 1, 033001. (<https://doi.org/10.1088/2632-2153/ab9a20>)

²⁹⁰ Raabe D, Mianroodi J, Neugebauer J. 2023 Accelerating the design of compositionally complex materials via physics-informed artificial intelligence. *Nature Computational Science*. 3, 198-209. (<https://doi.org/10.1038/s43588-023-00412-7>)

²⁹¹ Lyngby P, Sommer Thygesen K. 2022. Data-driven discovery of 2D materials by deep generative models. *npj Computational Materials*. 8, 232. (<https://doi.org/10.1038/s41524-022-00923-3>).

²⁹² Rao et. al. 2022 Machine learning-enabled high-entropy alloy discovery. *Science*. 378, 78-85. (<https://doi.org/10.1126/science.abo4940>).

²⁹³ Vasylenko et. al. 2021 Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. *Nature Communications*. 12, 5561. (<https://doi.org/10.1038/s41467-021-25343-7>).

²⁹⁴ Sun et. al. 2019. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances*. 5, 11. (<https://doi.org/10.1126/sciadv.aay4275>).

²⁹⁵ Stanev et. al. 2018. Machine learning modelling of superconducting critical temperature. *npj Computational*

Аутоматизовано експериментисање

Још једна важна примена вештачке интелигенције у материјалима је аутоматизовано експериментисање (АЕ). Централна идеја је интеграција вештачке интелигенције и роботике у затворену експерименталну петљу, где АЕ систем може да прође кроз итеративни експериментални процес заснован на претходном знању, побољшавајући резултат након сваке итерације²⁹⁶291. Постоји људска интервенција за покретање експеримента и дефинисање циљева.

Већа примена АЕ система у лабораторијама би имала значајан утицај на дизајн материјала и експериментисање. Прво, повећало би брзину истраживања уз смањене трошкове по експерименту²⁹⁷292. Ово би ослободило научнике да раде на другим задацима, уместо да се фокусирају на рутинске, дуготрајне експерименте, а такође би омогућило бржи развој и комерцијализацију технолошки релевантних материјала²⁹⁸293. Глобална мрежа интегрисаних АЕ система би повећала доступност истраживања у науци о материјалима, отварајући напредне технике истраживачким групама са мање ресурса. Иако АЕ технологије постоје и користе се за истраживање материјала, постоји неколико препрека које треба решити пре ширег усвајања. Прво, била би потребна побољшања експерименталног хардвера и софтвера, укључујући интерфејсе и алате који нису заштићени ауторским правима за ефикасну карактеризацију узорака. Сарадња између научника који се баве материјалима и стручњака за вештачку интелигенцију била би корисна, као и разумевање да ће људско учешће бити неопходно тамо где постоје безбедносна или етичка питања. Коначно, АЕ системима би био потребан приступ широком спектру података, укључујући метаподатке и негативне резултате, како би их обезбедили систему као претходно знање, што захтева стандардизацију управљања подацима и њиховог дељења.

Постоји неколико примера АЕ система који се успешно користе за истраживање материјала. Један такав пример је демонстрација Николајева и сарадника из 2016. године²⁹⁹294. о првој употреби Аутономног истраживачког система за оптимизацију раста угљеничних наноцеви, материјала који има потенцијалну употребу у технологијама хватања угљеника, као и запањујући низ тренутних примена³⁰⁰295. Међутим, угљеничне наноцеви су скупе; цена зависи од конфигурације и квалитета, али 1 грам се продаје од око 100 до више од 1200 фунти. АЕ се може користити за побољшање раста угљеничних наноцеви брзим понављањем параметара раста за оптимизацију својстава и веће приносе.

Materials. **4**, 29. (<https://doi.org/10.1038/s41524-018-0085-8>).

²⁹⁶ Stach *et. al.* 2021 Autonomous experimentation systems for materials development: A community perspective. Matter. **4**, 2702-2726. (<https://doi.org/10.1016/j.matt.2021.06.036>).

²⁹⁷ Stein H, Gregoire J. 2019 Progress and prospects for accelerating materials science with automated and autonomous workflows. Chemical Science. **10**, 9640. (<https://doi.org/10.1039%2Fc9sc03766g>)

²⁹⁸ Maruyama *et. al.* 2023 Artificial intelligence for materials research at extremes. MRS Bulletin. **47**, 1154-1164. (<https://doi.org/10.1557/s43577-022-00466-4>)

²⁹⁹ Nikolaev *et. al.* 2016 Autonomy in materials research: a case study in carbon nanotube growth. npj Computational Materials. **2**, 16031. (<https://doi.org/10.1038/npjcompumats.2016.31>).

³⁰⁰ De Volder M, Tawfik S, Baughman R, Hart J. 2013 Carbon Nanotubes: Present and Future Commercial Applications. Science. **339**, 535-539. (<https://doi.org/10.1126/science.1222453>).

ПОГЛАВЉЕ 4

ИСТРАЖИВАЊЕ, ИНОВАЦИЈЕ И ПРИВАТНИ СЕКТОР

ИСТРАЖИВАЊЕ, ИНОВАЦИЈЕ И ПРИВАТНИ СЕКТОР

Велика улагања приватног сектора у вештачку интелигенцију и њен значај у научним истраживањима имају различите импликације. Оне укључују централизацију критичне дигиталне инфраструктуре³⁰¹292; привлачење талената из академске заједнице у приватни сектор³⁰²293; и изазове отвореној науци³⁰³294.

Утицај приватног сектора на развој вештачке интелигенције за науку није необичан. Историјски гледано, аутоматизацију задатака покретали су актери из индустрије у тежњи ка смањењу трошкова рада и већој скалабилности³⁰⁴295. Данас, приватни сектор наставља да игра истакнуту улогу у унапређењу научних истраживања, при чему многе компаније имају научне програме вођене вештачком интелигенцијом, као што су Alphabet-ов Google DeepMind и Microsoft-ов AI for Science³⁰⁵296.

Улога приватног сектора у науци се такође шири, јер многе компаније доприносе обезбеђивању основних ресурса као што су рачунарска снага, приступ подацима и нове технологије вештачке интелигенције широј истраживачкој заједници³⁰⁶297.

Ово поглавље испитује растућу улогу приватног сектора у науци, ослањајући се на наручени преглед глобалног патентног пејзажа вештачке интелигенције, који описује расподелу власништва, развој и утицај технологија вештачке интелигенције. Такође прикупља перспективе из радионице о ризицима безбедности вештачке интелигенције и нарученог историјског прегледа.

³⁰¹ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

³⁰² Gofman M, Jin Z. 2022 Artificial Intelligence, Education, and Entrepreneurship. *Journal of Finance*, *Forthcoming*. (<https://doi.org/10.1111/jofi.13302>)

³⁰³ *Ibid*

³⁰⁴ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

³⁰⁵ Microsoft. Microsoft Research. AI4Science. See <https://www.microsoft.com/en-us/research/lab/microsoft-researchai4science/> (accessed 21 December 2023)

³⁰⁶ Kak A, Myers West S, Whittaker M. 2023 Opinion: Make no mistake – AI is owned by Big Tech. *MIT Technology Review*. See <https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/> (accessed 21 December 2023)

Променљиви пејзаж технологија вештачке интелигенције у научним истраживањима

Јавна и приватна улагања у вештачку интелигенцију за научни напредак се повећавају³⁰⁷298. На пример, влада Уједињеног Краљевства, преко UKRI-ја, издвојила је 54 милиона фунти универзитетима широм земље за подршку одговорном развоју вештачке интелигенције и финансирање истраживачких пројеката заснованих на вештачкој интелигенцији, а додатних 50 милиона фунти за убрзање истраживачких подухвата са индустријом и трећим сектором³⁰⁸299. Међутим, 2023. године, Microsoft је уложио преко 8 милијарди фунти у OpenAI³⁰⁹300, а Meta је обећала око 26,5 милијарди фунти за проширење својих „капацитета вештачке интелигенције“³¹⁰301. Процењује се да је 2022. године приватни сектор чинио 67% улагања у вештачку интелигенцију у ЕУ, док је јавни сектор допринео са 33%³¹¹302.

Један од начина разумевања променљивог пејзажа технологија вештачке интелигенције у научним истраживањима јесте посматрање трендова интелектуалне својине (IP). IP се односи на творевине ума, као што су проналасци, дизајни или писменост и уметничка дела. Закон о интелектуалној својини обухвата ауторска права, жигове, пословне тајне и патенте који дају ексклузивна права проналазачима или цесионарима на ограничено време у замену за јавно објављивање проналаска³¹²303. Овај одељак се првенствено ослања на трендове у патентима (видети Оквир 4) и бави се променљивим пејзажом технологија вештачке интелигенције у научним истраживањима.

BOX 4

IP Pragmatics 2023: Анализа глобалног патентног пејзажа

Краљевско друштво је наручило глобални преглед патентног пејзажа о патентима вештачке интелигенције за научна истраживања. Анализа је проценила власништво, развој и утицај патената вештачке интелигенције међу земљама, организацијама и индустријама у последњих 10 година. Такође је идентификовала кључне актере, трендове и потенцијалне импликације за научну заједницу.

Ова анализа је дефинисала вештачку интелигенцију као студију у рачунарству усмерену на развој машина и система који могу да обављају задатке за које се сматра да захтевају људску интелигенцију, као што су машинско учење или вештачке неуронске мреже. Кључни термини за претрагу и релевантни кодови Међународне класификације патената (Patent Classification (IPC))

³⁰⁷ IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society. See <https://royalsociety.org/newsresources/projects/science-in-the-age-of-ai/>

³⁰⁸ UK Government. £54 million boost to develop secure and trustworthy AI research. Gov.UK. See <https://www.gov.uk/government/news/54-million-boost-to-develop-secure-and-trustworthy-ai-research> (accessed 21 December 2023)

Bass D. 2023 Microsoft invests \$10 billion in ChatGPT maker OpenAI. *Bloomberg*. 23 January 2023. See <https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai> (accessed 21 December 2023). ³⁰⁹

³¹⁰ Targett E. 2023 Meta to spend up to \$33 billion on AI, as Zuckerberg pledges open approach to LLMs. *The Stack*. 27 April 2023. See <https://www.thestack.technology/meta-ai-investment/> (accessed 21 December 2023).

³¹¹ *Ibid*

³¹² Intellectual property and your work. Gov.UK. See: <https://www.gov.uk/intellectual-property-an-overview> (accessed March 22 2024)

коришћени су за свеобухватну идентификацију патената повезаних са вештачком интелигенцијом (видети извештај IP Pragmatics за више информација)³¹³304.

Иако прегледи патентног пејзажа могу пружити корисне налазе, постоје ограничења. Добијање комплетних глобалних података о патентима може бити изазовно, а анализа може представити непотпуну слику због ограничења у доступности или приступачности података. Иако се време кашњења може разликовати између јурисдикција за неколико месеци, постоји кашњење од 18 месеци између приоритетне пријаве и објављивања патента у већини главних глобалних територија. То значи да су подаци у овом прегледу од 2021. до 2023. непотпуни³¹⁴305.

Поред тога, сами подаци о патентима не обухватају пуни обим истраживања и развоја вештачке интелигенције, јер неке иновације можда нису патентирани или могу бити заштићено другим законима о интелектуалној својини, укључујући ауторска права, пословне тајне и заштитне знакове.

1. Раст у области патената за вештачку интелигенцију

Област патената за вештачку интелигенцију је нагло порасла у последњој деценији, са приближно 74% укупног броја поднетих патената у последњих пет година³¹⁵306 (видети слику 2 на страни 67 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf). У 2022. години, тржишна вредност је достигла 109,718 милијарди фунти, са пројектованом сложеностом годишњом стопом раста (CAGR) од 37,3% од 2023. до 2030. године³¹⁶307. Приметно је да Кина предњачи у подношењу патената за вештачку интелигенцију, држећи приближно 62% тржишта, а следе је Сједињене Државе са око 13,2%³¹⁷308. Пројектовано је да ће Азијско-пацифички регион имати највишу CAGR (48,6%) између 2022. и 2027. године, у поређењу са САД (43%) и Европом (46,5%). Овај налаз допуњује пријављени пораст иновација вештачке интелигенције у здравственој индустрији у Кини и Индији^{318,319}309,310

2. Глобални тржишни удео у вештачкој интелигенцији за науку

Употреба вештачке интелигенције на тржишту науке и инжењерства првенствено је вођена потражњом за технологијом вештачке интелигенције ради покретања иновација и економског раста. Стога постоји корелација између трендова подношења патентних захтева и глобалног тржишног удела³²⁰311. Северна Америка, са својом богатом концентрацијом технолошких фирми и

³¹³ IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society.

See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

³¹⁴ *Ibid.*

³¹⁵ *Ibid*

³¹⁶ Grand View Research. See <https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market#> (Accessed 21 December 2023)

³¹⁷ IP Pragmatics, 2024 Artificial intelligence related inventions. The Royal Society.

See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

³¹⁸ Nair M, Sethumadhaven A. 2022 AI in healthcare: India's trillion-dollar opportunity. *World Economic Forum*. See <https://www.weforum.org/agenda/2022/10/ai-in-healthcare-india-trillion-dollar> (accessed 21 December 2023)

³¹⁹ Olcott E. 2022, China sets the pace in adoption of AI in healthcare technology. *Financial Times*. 31 January 2022. See: <https://www.ft.com/content/c1fe6fbf-8a87-4328-9e75-816009a07a59> (accessed 21 December 2023)

³²⁰ BBC Research. 2022 Global Markets for Machine Learning in the Life Sciences. October 2022. See <https://www.bccresearch.com/market-research/healthcare/global-markets-for-machine-learning-in-life-sciences.html> (accessed 21 December 2023)

квалификованих стручњака, доминира овим тржиштем³²¹312. У Европи, Немачка предњачи, али се Уједињено Краљевство истиче са значајним уделом од 14,7% на тржишту вештачке интелигенције за науку о животу и има највишу прогнозирану сложену годишњу стопу раста (CAGR) у региону од 47,9%³²²313

Глобална дистрибуција броја патентних породица везаних за вештачку интелигенцију по земљи првог приоритета погледајте на слици 3, страна 68-69 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Велика Британија, која је на 10. месту у свету и 2. у Европи по броју поднетих патената, показује снажан потенцијал раста³²³314. Канцеларија за интелектуалну својину Велике Британије (UKIPO) усваја приступ који је више прилагођен носиоцима патената приликом испитивања проналазака имплементираних рачунаром и вештачке интелигенције у поређењу са Европским патентним заводом (ЕРО), што је наглашено недавним одлукама попут случаја „Емоционална перцепција вештачке интелигенције против Генералног контролора патената“³²⁴315.

Овај случај је довео до прилагођавања праксе испитивања UKIPO-а, уклањајући специфичне смернице о вештачким неуронским мрежама³²⁵316. Иако је на ову одлуку уложена жалба и тренутно чека на преиспитивање, недавне пресуде су ојачале позицију Велике Британије као преферираног региона за заштиту интелектуалне својине везане за вештачку интелигенцију, јачајући њену улогу кључног играча у иновацијама вештачке интелигенције.

Глобални тржишни удео машинског учења у природним наукама, по регионима, 2021. (%) погледајте на слици 4, страна 70 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Међутим, глобални пејзаж је обележен разликама. Скупи и сложени процеси пријаве патената, посебно у регионима попут Африке, представљају значајне препреке. На пример, патентирање у Африци преко Афричке регионалне организације за интелектуалну својину кошта преко 29.000 фунти, што је знатно више него у Великој Британији, где је цена око 1.900 фунти.

Упркос порасту броја афричких технолошких центара, високи трошкови регистрације интелектуалне својине и недостатак јединственог система отежавају патентирање³²⁶317. Иницијативе попут

³²¹ *ibid*

³²² *ibid*

³²³ IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

³²⁴ Emotional perception ai ltd v comptroller-general of patents, designs and trademarks. 2023. Find case law – The National Archives. See <https://caselaw.nationalarchives.gov.uk/ewhc/ch/2023/29482948> (accessed 4 March 2024).

³²⁵ Examination of patent applications involving Artificial Neural Networks (ANN). Gov.UK. See <https://www.gov.uk/government/publications/examination-of-patent-applications-involving-artificial-neuralnetworks/examination-of-patent-applications-involving-artificial-neural-networks-ann> (accessed 4 March 2024).

³²⁶ Lewis J, Schneegans S, Straza T. 2021 UNESCO Science Report: The race against time for smarter *development*. UNESCO Publishing. See: <https://unesdoc.unesco.org/ark:/48223/pf0000377250> (accessed 22 March 2024)

Панафричке организације за интелектуалну својину имају за циљ да се позабаве овим изазовима, иако се тренутно суочавају са оперативним кашњењима³²⁷318

Европски тржишни удео машинског учења у природним наукама, по земљама, 2021. (%) погледајте на слици 5, страна 71 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

3. Кључни играчи у вештачкој интелигенцији за науку патентни пејзаж

Што се тиче технолошког утицаја (означеног бројем пута колико се патент цитира у каснијем патенту или унапред наведеним цитатима), САД се истичу по томе што имају вредне патенте. Упоредно, упркос значајном расту Индије у броју поднетих патената за вештачку интелигенцију, она још увек није постигла велики технолошки утицај. Док Велика Британија, иако представља мањи део патентног пејзажа, показује утицај на истраживање и иновације, рангирајући се међу највећим у свету³²⁸319.

Анализа 20 највећих носилаца патената везаних за вештачку интелигенцију наглашава активно учешће и индустрије и академских субјеката у широј научној и инжењерској истраживачкој сфери. Приметно је да су се компаније попут Canon-а, Alphabet-а, Siemens-а, IBM-а и Samsung-а појавиле као кључни доприносиоци, са значајним портфолијима патената који имају значајан утицај у научним и инжењерским областима. Упркос доминацији комерцијалних субјеката у већини региона, академске институције, укључујући Универзитет у Оксфорду, Империјал колеџ у Лондону и Универзитет у Кембриџу, истакнуто се налазе међу водећим подносиоцима патената у Великој Британији³²⁹320, што указује на мешавину сарадње између академије и индустрије и независних доприноса³³⁰321.

Изазови везани за улогу приватног сектора у науци заснованој на вештачкој интелигенцији

Поред разматрања трендова патентирања, Краљевско друштво је истражило изазове учешћа приватног сектора у научним истраживањима заснованим на вештачкој интелигенцији. Уочи Глобалног самита о безбедности вештачке интелигенције, који је Уједињено Краљевство организовало 2023. године, Краљевско друштво и Министарство за науку, иновације и технологију Уједињеног Краљевства (DSIT) сазвали су радионицу о безбедносним ризицима вештачке интелигенције у научним истраживањима³³¹322. Идентификовани изазови укључују:

1. Доминација приватног сектора и централизација развоја науке засноване на вештачкој интелигенцији

³²⁷ *Ibid.*

³²⁸ IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

³²⁹ *Ibid*

³³⁰ Legislation.Gov.UK. Copyright, Designs and Patents Act 1988. See: <https://www.legislation.gov.uk/ukpga/1988/48/contents>

³³¹ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/> (accessed 7 May 2024)

Централизација развоја вештачке интелигенције под великим технолошким фирмама (нпр. Google, Microsoft, Amazon, Meta и Alibaba) могла би довести до корпоративне доминације над инфраструктуром кључном за научни напредак. То укључује власништво над огромним скуповима података за обуку модела вештачке интелигенције, огромним рачунарским инфраструктурама и врхунским талентима за вештачку интелигенцију³³²323.

Централизација може ограничити шире учешће у управљању агендом истраживања вештачке интелигенције и може довести до малог броја доносилаца одлука који обликују која истраживања се спроводе и објављују из утицајних индустријских лабораторија. На пример, контроверзно отпуштање истраживача вештачке интелигенције Dr Timnit Gebru из компаније Google истакло је непрозирно интерно доношење одлука у истраживачким јединицама приватног сектора.

Ово је пример неусклађености између корпоративног интереса да заштите своју тржишну предност и академских вредности да отворено испитују друштвени утицај унапређења технологије³³³324.

2. Прекомерно ослањање на алате које покреће индустрија и референтне вредности за науку засновану на вештачкој интелигенцији

Растући утицај индустрије у постављању кључних референтних вредности³³⁴325, развоју најсавременијих модела³³⁵326 и управљању академским публикацијама³³⁶327 централизује контролу над екосистемима вештачке интелигенције, обухватајући хардвер, софтвер и податке³³⁷328

Патентни систем има значајан утицај на тржишта, стимулишући конкуренцију и раст у свим секторима. Кроз овај систем, власници патената могу да обезбеде повраћај инвестиција и контролу ширења технологије. Међутим, без надзора постоји ризик да ће ови ентитети дати приоритет комерцијалним интересима у односу на шири научни напредак, контролишући критичну инфраструктуру, базе података и алгоритме како би одржали своју доминацију на тржишту³³⁸329.

³³² Kak A, Myers West S, Whittaker M. 2023 Opinion: Make no mistake – AI is owned by Big Tech. *MIT Technology Review*. 5 December 2023 See <https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/>. (accessed 21 December 2023)

³³³ Hao K. 2020 We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review*. 4 December 2020. See <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-researchpaper-forced-out-timnit-gebru/> (accessed 12 July 2023)

³³⁴ Hodak, M., Ellison, D., & Dholakia, A. (2020, August). Benchmarking AI inference: where we are in 2020. In *Technology Conference on Performance Evaluation and Benchmarking* (pp. 93-102). Cham: Springer International Publishing

³³⁵ *Ibid*

³³⁶ Ahmed N, Wahed M, Thompson, N. C. 2023. The growing influence of industry in AI research. *Science*, 379(6635),884-886. (<https://doi.org/10.1126/science.ade2420>)

³³⁷ The Royal Society interviews with scientists and researchers. 2022 - 2023

³³⁸ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

Концентрација моћи и ресурса не само да би могла да омета конкуренцију и транспарентност, већ и да успостави јединствене тачке отказа, што изазива забринутост у вези са отпорношћу и отвореношћу научних истраживања заснованих на вештачкој интелигенцији³³⁹330.

BOX 5

Улога приватног сектора у патентирању лекова и фармацеутских изума

Анализа патената вештачке интелигенције у медицини и фармацеутским интервенцијама показује да, иако су Универзитет Харвард и Масачусетски технолошки институт (МИТ) били пионири у овој области, портфолији патената које поседују Roche и IBM изгледају највреднији у овом сектору³⁴⁰331. Alphabet је проширио свој утицај кроз подружнице попут Google DeepMind, који је развио AlphaFold– систем вештачке интелигенције који је револуционисао предвиђање структуре протеина³⁴¹332. Ова иновација означава значајан помак у пејзажу медицинских патената^{342,343}333,334, подстичући друге технолошке гиганте попут Мајкрософта да инвестирају у сличне технологије. Ово истиче комерцијални потенцијал и конкурентску динамику у овом сектору³⁴⁴335.

Међутим, Alphabet-ова стратегија подношења кластера патената, која се сматра „ограђивањем патената“, сугерише шири тренд фирми које користе своје интелектуално право како би заштитиле и прошириле своју тржишну позицију³⁴⁵336. Такви приступи не само да штите од кршења, већ и спречавају конкуренте да развијају сродне технологије, јачајући Alphabet-ову – а самим тим и Google DeepMind-ову – доминацију у медицинским истраживањима вођеним вештачком интелигенцијом

3. Приватни сектор и отворена наука

Комерцијални подстицаји који подстичу приватно власништво над подацима за истраживања заснована на вештачкој интелигенцији могли би да ограниче праксе отворене науке. Ово ограничава способност научника ван индустрије да праведно доприносе и испитују податке за системе вештачке интелигенције заједно са колегама из индустрије.

Подаци у приватном власништву су често комерцијално осетљиви и могу захтевати споразуме о неоткривању, што потенцијално утиче на интегритет истраживања. Подаци који се у почетку

³³⁹Westgarth, T., Chen, W., Hay, G., & Heath, R. 2022 Understanding UK Artificial Intelligence R&D commercialization and the role of standards. See https://oxfordinsights.com/wp-content/uploads/2023/10/DCMS_and_OAI_-_Understanding_UK_Artificial_Intelligence_R_D_commercialisation_accessible-1.pdf (accessed 21 December 2023)

³⁴⁰ IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See <https://royalsociety.org/newsresources/projects/science-in-the-age-of-ai/>

³⁴¹ Google DeepMind. Technology: AlphaFold. See <https://deepmind.google/technologies/alphafold/> (accessed 21 December 2023)

³⁴² Borkakoti N, Thornton J.M., 2023. AlphaFold2 protein structure prediction: Implications for drug discovery. *Current opinion in structural biology*, **78**, p.102526 (<https://doi.org/10.1016/j.sbi.2022.102526>)

³⁴³ Jumper, J., *et al.* 2021 Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873), pp.583-589.(doi: 10.1038/s41586-021-03819-2)

³⁴⁴ IP Pragmatics. 2024 Artificial intelligence related inventions. The Royal Society. See <https://royalsociety.org/newsresources/projects/science-in-the-age-of-ai/>

³⁴⁵*Ibid*

сматрају ниским ризиком могу касније добити комерцијалну вредност и бити повучени, као што се види код неких компанија друштвених медија које поощтравају приступ подацима након пораста обуке за мастер студије права о јавним подацима^{346,347}337,338.

Алтернативни приступи монетизацији, попут подстицања лиценцирања језера података (GRUBI PODACI U PRIRODNOM OBLIKU) и коришћења одредби база података, могу пружити отворенији и прагматичнији приступ дељењу података³⁴⁸339.

Даљи приступи укључују промене закона, као што су захтеви за компаније друштвених медија да деле податке у Европском закону о дигиталним услугама³⁴⁹340 и принципи интервенције за откључавање вредности података у целој економији у Националној стратегији података Уједињеног Краљевства³⁵⁰341. Поред тога, технички приступи укључују технологије за побољшање приватности³⁵¹342 и законе о сајбер безбедности како би се обезбедиле правне мере и осигурао безбеднији хардвер и софтвер³⁵²343. Отворени код и платформе такође нуде неке предности организацијама приватног сектора, укључујући брзину и исплативост, али имају и значајна ограничења, укључујући недостатак подршке, безбедносне ризике и компатибилност. На пример, индустријска партнерства ради обостране користи, као што је партнерство између Сименса и Мајкрософта, могу подстаћи усвајање вештачке интелигенције у различитим индустријама дељењем софтвера, хардвера и талената³⁵³344. Током пандемије COVID-19, неке приватне организације су се одрекле патентних права за опште добро, а водеће технолошке компаније су донирале своје патенте иницијативама отвореног кода³⁵⁴345.

4. Улога приватног сектора у безбедности вештачке интелигенције

Доминација приватног сектора у вештачкој интелигенцији за науку такође представља изазове за безбедност вештачке интелигенције. Организације и институције које воде развој вештачке интелигенције често користе сопствену способност да процене штету, успоставе заштитне мере и безбедно објаве своје моделе. Као што је описано од стране OpenAI у раду који стоји иза

³⁴⁶Isaac M. 2023 Reddit wants to get paid for helping to teach big AI systems. *The New York Times*. 18 April 2023. See <https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html> (accessed 21 December 2023).

³⁴⁷Murphy H. 2023 Elon Musk rolls out paywall for Twitter's data. *The Financial Times*. 29 April 2023. See <https://www.ft.com/content/574a9f82-580c-4690-be35-37130fba2711> (accessed 21 December 2023).

³⁴⁸Grossman R L. 2019 Data lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data. *Trends in Genetics*, 35(3), pp.223-234. (<https://doi.org/10.1016/j.tig.2018.12.006>)

³⁴⁹European Commission. The Digital Services Act. See https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (accessed 5 February 2024)

³⁵⁰Department for Science, Innovation and Technology. National Data Strategy. 5 December 2022. See <https://www.gov.uk/guidance/national-data-strategy> (accessed 5 February 2024)

³⁵¹The Royal Society. 2023 From privacy to partnership. See <https://royalsociety.org/topics-policy/projects/privacyenhancing-technologies/> (accessed 21 December 2023)

³⁵²European Commission. Directive on measures for a high common level of cybersecurity across the Union (NIS2 Directive). See <https://digital-strategy.ec.europa.eu/en/policies/nis2-directive> (accessed 22 February 2024)

³⁵³Siemens. 2024 Siemens and Microsoft partner to drive cross-industry AI adoption. See <https://press.siemens.com/global/en/pressrelease/siemens-and-microsoft-partner-drive-cross-industry-ai-adoption> (accessed 26 February 2024)

³⁵⁴UNESCO Recommendation on Open Science. 2021. See: <https://www.unesco.org/en/legal-affairs/recommendationopen-science> (accessed 6 February 2024)

објављивања GPT-4, комерцијални подстицаји и безбедносна разматрања могу доћи у сукоб са научним вредностима као што су транспарентност и праксе отворене науке³⁵⁵346.

Hugging Face, организација отвореног кода, предлаже процену компромиса за безбедно и одговорно објављивање, као што је илустровано у Gradient of System Access³⁵⁶347 (видети слику 6). Сличне оквири могу размотрити и развити научне заједнице како би процениле услове под којима је објављивање података за обуку безбедно, што им омогућава да допринесу научном напретку, а истовремено смањују потенцијал за штету и злоупотребу.

Универзитети такође могу играти кључну улогу у унапређењу безбедности вештачке интелигенције, промовишући етичке истраживачке стандарде или подстичући академска истраживања о штетности вештачке интелигенције. Међутим, они немају исте могућности као велике технолошке компаније да уведу робусне заштитне мере и најбоље праксе у свим аспектима развоја сложене вештачке интелигенције. У последње време, националне владе придају већи значај дискусијама о безбедности вештачке интелигенције. Од Глобалног самита о безбедности вештачке интелигенције у новембру 2023. године, Велика Британија је покренула Институт за безбедност вештачке интелигенције³⁵⁷348, док су САД најавиле оснивање Америчког института за безбедност вештачке интелигенције у оквиру Националног института за стандарде и технологију (NIST)³⁵⁸349.

Репродукција градијента системског приступа коју је развио Hugging Face ³⁵⁹350 погледајте на слици 6, страна 77 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

„Слобода, иновативност и креативност академске заједнице уз ресурсе, структуру и управљање приватног сектора... то је било потпуно ослобађајуће.“Учесник интервјуа Краљевског друштва говори о заједничким академским и индустријским улогама

Могућности за међусекторску сарадњу

Међусекторска сарадња нуди значајне могућности, користећи иновативне и образовне снаге академске заједнице са ресурсима и практичним фокусом индустрије³⁶⁰351. Упркос забринутости због тога што систем патената централизује развој вештачке интелигенције, она

³⁵⁵ OpenAI et al. 2023 Gpt-4 technical report. arxiv 2303.08774. View in Article, 2, p.13. (<https://doi.org/10.48550/arXiv.2303.08774>)

³⁵⁶ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

³⁵⁷ Gov.UK. 2024 Introducing the AI Safety Institute. See <https://www.gov.uk/government/publications/ai-safety-institute/overview/introducing-the-ai-safety-institute> (accessed 26 February 2024)

³⁵⁸ NIST. 2024 U.S. Artificial Intelligence Safety Institute. See <https://www.nist.gov/artificial-intelligence/artificialintelligence-safety-institute> (accessed 26 February 2024)

³⁵⁹ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

³⁶⁰ Wright B et al. 2014 Technology transfer: Industry-funded academic inventions boost innovation. *Nature* **507**, 297–299. <https://doi.org/10.1038/507297a>

такође може да подстакне сарадњу. Објављене пријаве патената повећавају технолошку транспарентност и обезбеђују ток прихода који може да подржи заједничка улагања између универзитета и индустрије.

Међутим, све веће присуство приватног сектора у финансирању науке засноване на вештачкој интелигенцији изазива забринутост да би утицај индустрије могао да помери фокус са фундаменталних истраживања на примењену науку³⁶¹352. Ова промена би могла да погорша „одлив мозга“³⁶²353, где значајан број талената у области вештачке интелигенције напушта академску заједницу и одлази у приватни сектор³⁶³354, вођен вишим платама, напредним ресурсима и могућношћу рада на практичним применама³⁶⁴355

Да би се супротставиле овом тренду, иницијативе попут Фонда за иновативну производњу у области природних наука у Великој Британији³⁶⁵356 (који укључује 17 милиона фунти државног финансирања и приватна улагања од 260 милиона фунти), показују како владине и приватне инвестиције могу синергијски подржати пројекте који покрећу иновације и економски раст³⁶⁶357. Овај модел сарадње не само да подстиче технолошки напредак, већ нуди и платформу академској заједници да се укључи у најсавременија истраживања, а истовремено користи ресурсе индустрије. Друга партнерства би могла да се прошире изван финансијских аспеката, обухватајући заједничке истраживачке пројекте³⁶⁷358, дељене публикације и интелектуалне размене на конференцијама или путем неформалних мрежа³⁶⁸359. Она такође нуде практичне могућности ангажовања попут пракси и одсуства, омогућавајући академицима да стекну искуство у индустрији без одступања од својих академских улога³⁶⁹360. Да би се оптимизовале предности међусекторске сарадње, универзитети и истраживачке институције такође могу развити робусне политике интелектуалне својине, подстичући окружење у којем су иновације заштићене и могу се ефикасно делити са индустријским партнерима. Стварањем структурираних путева за сарадњу, као што су заједнички напори у вези са патентирањем или уговори о лиценцирању, оба сектора могу допринети унапређењу истраживања вештачке интелигенције, истовремено решавајући изазове попут разлика у ресурсима и приватности података.

³⁶¹*bid*

³⁶² Kunze L. 2019. Can we stop the academic AI brain drain? *KI-Künstliche Intelligenz*, 33(1), 1-3.

(<https://doi.org/10.1007/s13218-019-00577-2>)

³⁶³ Gofman M, Jin Z. 2022 Artificial Intelligence, Education, and Entrepreneurship. *Journal of Finance, Forthcoming*.

(<https://doi.org/10.1111/jofi.13302>)

³⁶⁴ UK universities alarmed by poaching of top computer science brains. *Financial Times*. 9 May 2018.

See <https://www.ft.com/content/895caede-4fad-11e8-a7a9-37318e776bab> (accessed 10 June 2023)

³⁶⁵ Life sciences companies supercharged with £277 million in government and private investment. Gov.UK

See <https://www.gov.uk/government/news/life-sciences-companies-supercharged-with-277-million-in-government-and-private-investment> (accessed 26 February 2024)

³⁶⁶ Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI. Gov.UK

See <https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai> (accessed 26 February 2023)

³⁶⁷ Evans JA (2010) Industry induces academic science to know less about more. *Am J Sociol* 116(2):389–452

³⁶⁸ Perkmann M, Walsh K (2007) University–industry relationships and open innovation: towards a research agenda.

Int J Manag Rev 9(4):259–280 (<https://doi.org/10.1111/j.1468-2370.2007.00225.x>)

³⁶⁹ Cohen WM, Nelson RR, Walsh JP. 2002 Links and impacts: the influence of public research on industrial R&D. *Manag Sci* 48(1):1–23. (<https://doi.org/10.1287/mnsc.48.1.1.14273>)

ПОГЛАВЉЕ 5

Истраживачка етика и безбедност вештачке интелигенције

„Можда нећемо желети да учинимо неке од скупова података доступним због лакоће злоупотребе. А то делује помало супротно од онога чему тежимо као научници. Али можда бисмо сада желели више да размишљамо о безбедности.“

Краљевско друштво

Истраживачка етика и безбедност вештачке интелигенције

Како се употреба вештачке интелигенције шири у научним дисциплинама, јављају се нови етички изазови око ненамерне или намерне злоупотребе вештачке интелигенције³⁷⁰364. Такође постоји све већа забринутост јавности у вези са праведном и етичном употребом њихових података³⁷¹365 и степеном у којем алати засновани на вештачкој интелигенцији могу ширити штетне предрасуде, дискриминацију и друштвене штете³⁷²366. Ризике безбедности вештачке интелигенције такође треба узети у обзир, јер је постало лакше пренаменити алгоритме за злонамерну употребу³⁷³367.

Уочи Глобалног самита о безбедности вештачке интелигенције, који је Уједињено Краљевство организовало 2023. године, Краљевско друштво и Министарство за науку, иновације и технологију Уједињеног Краљевства (DSIT) сазвали су радионицу о безбедности вештачке интелигенције. Следеће теме су се појавиле као етички изазови повезани са употребом вештачке интелигенције у научним истраживањима:

1. Подаци и алгоритамска пристрасност

Системи вештачке интелигенције могу имати уграђене пристрасности кроз податке за обуку и алгоритамски дизајн. Када се не ублаже, алгоритамска пристрасност може довести до неправедних исхода и погоршати неједнакости³⁷⁴368. Интеграција вештачке интелигенције у медицину је, на пример, истакла како алгоритамске пристрасности могу довести до нетачних медицинских дијагноза, неадекватног лечења и погоршања разлика у здравственој заштити^{375,376}369,370. Ако се пристрасност података пренесе на податке за обуку модела вештачке интелигенције, постоји ризик да се модели неће добро уклопити у друге заједнице³⁷⁷371 (видети Оквир 6).

³⁷⁰Ghotbi N. 2024. Ethics of Artificial Intelligence in Academic Research and Education. In *Second Handbook of Academic Integrity* (pp. 1355-1366). (https://doi.org/10.1007/978-981-287-079-7_143-1)

³⁷¹Lomas N. 2023 UK court tosses class-action style health data misuse claim against Google Deepmind. Tech Crunch. 19 May 2023. See <https://techcrunch.com/2023/05/19/uk-court-tosses-class-action-style-health-data-misuse-claimagainst-google-deepmind> (accessed 21 December 2023)

³⁷²Brennan, J. 2023. AI assurance? Assessing and mitigating risks across the AI lifecycle. Ada Lovelace Institute. See <https://www.adalovelaceinstitute.org/report/risks-ai-systems/> (accessed September 30 2023)

³⁷³Urbina F, Lentzos F, Invernizzi C, Ekins S. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3), 189-191. (<https://doi.org/10.1038/s42256-022-00465-9>)

³⁷⁴UK Parliament POST. 2024. Policy implications of artificial intelligence (AI). <https://researchbriefings.files.parliament.uk/documents/POST-PN-0708/POST-PN-0708.pdf>

³⁷⁵Panch, T., Mattie, H. and Atun, R. 2019. Artificial intelligence and algorithmic bias: implications for health systems. *Journal of global health*, 9(2).

³⁷⁶Celi, L.A et al., 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, 1(3), p.e0000022.

³⁷⁷Royal Society and Department for Science, Innovation, and Technology workshop on horizon-scanning AI safety risks across scientific disciplines, 2023.

Алгоритамске штете се такође могу манифестовати у области финансирања и научне комуникације. Алати вештачке интелигенције се користе да би почетне провере процеса доделе грантова и рецензирања биле мање временски захтевне³⁷⁸372. Између осталих примена, могу помоћи рецензентима у идентификовању лажних цитата³⁷⁹373, побољшању квалитета радова³⁸⁰374 и смањењу плагијата³⁸¹375

Иако вештачка интелигенција може допринети смањењу пристрасности „првог утиска“ код људских рецензента, она такође може појачати већ постојеће родне, језичке или институционалне предрасуде укорењене у скуповима података за обуку³⁸²376, што може наштетити могућностима за напредовање у каријери за недовољно заступљене научнике³⁸³377. Иницијативе попут пројекта GRAIL истражују етичке принципе и најбоље праксе за коришћење вештачке интелигенције у финансирању и евалуацији истраживања³⁸⁴378

BOX 6

Вишејезични језички модели

Развој вишејезичних језичких модела може допринети смањењу пристрасних резултата. На пример, BLOOM (BigScience Language Open-science Openaccess Multilingual) је тренутно највећи модел отвореног истраживачког језика развијен са циљем смањења штетних и пристрасних резултата обучавањем на мањем избору квалитетнијих, вишејезичних текстуалних извора. BLOOM је развило 1000 истраживача и обучен је на 46 различитих језика и 13 програмских језика³⁸⁵379. Подаци о обуци су доступни, а модел је развијен у складу са етичком повељом која у средиште ставља вредности као што су разноликост, инклузивност, репродуктивност; и има за циљ неговање приступачности, вишејезичности и интердисциплинарности³⁸⁶380.

2. Халуцинације и дезинформације генерисане вештачком интелигенцијом

Растућа употреба општих или основних модела у науци (нпр. генеративна вештачка интелигенција и LLM) доводи до јединствених разматрања у вези са етиком и безбедношћу. На пример, док се LLM-ови могу користити за убрзавање академског писања, они се такође могу користити за намерно генерисање научних дезинформација³⁸⁷381. Повећан јавни приступ LLM-овима смањује препреке за злонамерне актере да генеришу убедљив садржај креиран од стране

³⁷⁸ Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. 2021 AI-assisted peer review. *Humanities and Social Sciences Communications*, **8**(1), pp.1-11. (<https://doi.org/10.1057/s41599-020-00703-8>)

³⁷⁹ The Royal Society roundtable on Large Language Models, July 2023

³⁸⁰ Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. 2021 AI-assisted peer review. *Humanities and Social Sciences Communications*, **8**(1), pp.1-11. (<https://doi.org/10.1057/s41599-020-00703-8>)

³⁸¹ Heaven D. 2022 AI peer reviewers unleashed to ease publishing grind. *Nature*. 22 November 2018.

See <https://www.nature.com/articles/d41586-018-07245-9> (accessed 27 March 2023)

³⁸² Checco A, Bracciale L, Loreti P, Pinfield S, Bianchi G. 2021 AI-assisted peer review. *Humanities and Social Sciences Communications*, **8**(1), pp.1-11. (<https://doi.org/10.1057/s41599-020-00703-8>)

³⁸³ Chawla, D.S. 2022, Should AI have a role in assessing research quality?. *Nature*. (DOI: 10.1038/d41586-022-03294-3)

³⁸⁴ Research on Research Institute. See <https://researchonresearch.org/project/grail/> (accessed 5 January 2024)

³⁸⁵ Hugging Face. Documentation of BLOOM. See: https://huggingface.co/docs/transformers/model_doc/bloom (accessed 21 December 2023)

³⁸⁶ Hugging Face. BigScience Ethical Charter. See: <https://bigscience.huggingface.co/blog/bigscience-ethical-charter> (accessed 21 December 2023)

³⁸⁷ Wang H *et al.* 2023 Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (<https://doi.org/10.1038/s41586-023-06221-2>)

машина, што смањује вероватноћу људског откривања³⁸⁸382 (видети Оквир 1 о вежби црвеног тима Краљевског друштва о научним дезинформацијама).

Употреба LLM-ова у научном пројекту такође може повећати изложеност „халуцинацијама“ – што се односи на генерисање убедљивих и реалистичних резултата који не одговарају улазима из стварног света. Чак и када нема злонамерне намере, технологије општег претходно обученог трансформатора (GPT) могу да фабрикују чињенице, податке и цитате када одговарају на упит. Брзи пораст машински генерисаних дезинформација на мрежи повећава ризик да ће следећа генерација модела обучених на подацима са веба деградирати у перформансама и апсорбовати изобличења и нетачности пронађене у измишљеном тексту и подацима³⁸⁹383

На пример, Метин(Meta’s LLM) мастер програм за науку, Галактика(Galactica), обучен је на 48 милиона научних чланака, веб-сајтова, уџбеника и других улазних података како би помогао истраживачима да сумирају литературу, генеришу академске радове, пишу научни код и аотирају податке (нпр. молекуле и протеине). Међутим, демо (, the demo) је паузиран након три дана коришћења. Један од највећих ризика које представља Галактика био је колико је самоуверено произвођила лажне информације и недостатак смерница за њихову идентификацију³⁹⁰384.

Као и код других облика дезинформација, халуцинације могу нарушити поверење јавности у науку³⁹¹385. Методе за валидацију и откривање вештачке интелигенције, као што су водени жигови или технологије порекла садржаја³⁹²386, истражују се како би се омогућило откривање садржаја генерисаног вештачком интелигенцијом и ублажиле потенцијалне штете узроковане халуцинацијама³⁹³387, као и да би се осигурало поверење јавности у нове системе вештачке интелигенције³⁹⁴388.

3. Двострука употреба технологија вештачке интелигенције развијених за науку

Двострука употреба система вештачке интелигенције односи се на ситуације у којима се систем развијен за одређену употребу затим присваја или модификује за другу употребу. Злонамерна употреба односи се на примене у којима је намера да се нанесе штета³⁹⁵389. Међу најистакнутијим и најдокументованијим примерима злонамерне употребе вештачке интелигенције јесте развој хемијског и биолошког оружја коришћењем система вештачке интелигенције који имају корисне примене за научна истраживања.

³⁸⁸ Bommasani *et al.* 2021. On the opportunities and risks of foundation models. See: <https://crfm.stanford.edu/assets/report.pdf> (accessed March 21 2024)

³⁸⁹ Pan Y, Pan L, Chen W, Nakov P, Kan M Y, Wang W Y. 2023. On the Risk of Misinformation Pollution with Large Language Models. *arXiv preprint (arXiv:2305.13661)*

³⁹⁰ MIT Review. Why Meta’s latest large language model survived only three days online 2022. See: <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-threedays-gpt-3-science/> (accessed September 30 2023)

³⁹¹ Bontridder, N. and Pouillet, Y., 2021. The role of artificial intelligence in disinformation. *Data & Policy*, 3, p.e32. (doi:10.1017/dap.2021.20)

³⁹² The Royal Society. Generative AI, content provenance and a public service internet. See: <https://royalsociety.org/news-resources/publications/2023/digital-content-provenance-bbc/>

³⁹³ Watermarking refers to techniques that can embed identification information into the original data, model, or content to indicate provenance or ownership.

³⁹⁴ Partnership on AI. PAI’s Responsible Practices for Synthetic Media. See: https://syntheticmedia.partnershiponai.org/#read_the_framework (accessed 21 December 2023)

³⁹⁵ Ueno, H. 2023. Artificial Intelligence as Dual-Use Technology. In *Fusion of Machine Learning Paradigms: Theory and Applications* (pp. 7-32). Cham: Springer International Publishing. (https://doi.org/10.1007/978-3-031-22371-6_2)

Године 2020, компанија Collaborations Pharmaceuticals, биофармацеутска компанија која гради моделе машинског учења како би помогла у откривању лекова и лечењу ретких болести, објавила је резултате о ономе што су назвали „тренутком у ком научимо нешто неочекивано“ (‘teachable moment’) у вези са употребом метода откривања лекова заснованих на вештачкој интелигенцији. Након позива Швајцарског федералног института за NBC (нуклеарну, биолошку и хемијску) заштиту, компанија је обучила генератор молекула заснован на вештачкој интелигенцији који се користи за откривање лекова да генерише токсичне молекуле унутар одређеног прага токсичности³⁹⁶390. Црпећи податке из јавне базе података, модел је за мање од 6 сати генерисао 40.000 молекула. Многи од ових молекула били су слични или токсичнији од нервног агенса VX, забрањеног и високо токсичног смртоносног хемијског оружја.

Иако теоретско стварање токсичних молекула не подразумева да је њихова производња одржива или изводљива, експеримент показује како вештачка интелигенција може убрзати процес стварања опасних супстанци, укључујући смртоносно биолошко оружје³⁹⁷391. Компанија је позвала на неколико акција за решавање безбедносних ризика, као и за праћење, решавање и ограничавање злонамерних апликација модела вештачке интелигенције који се користе у науци³⁹⁸392.

4. Тровање података и напади машинског учења

Мере за побољшање безбедности вештачке интелигенције у науци морају да узму у обзир развој робусних модела који могу да издрже нападе засноване на подацима³⁹⁹393. Обука модела вештачке интелигенције на огромним и лоше курираним скуповима података ствара рањивости на случајеве „тровања података“, „убризгавања лажних података“ или напада (one-pixel attack) „једног пиксела“⁴⁰⁰394. Ове тактике укључују уметање бучних, нетачних или манипулисаних података како би се обманули системи машинског учења, а да се притом људи не примете или тешко открију^{401,402,403}395,396,397. „Отровани“ или манипулисани скупови података један су од најчешћих и документованих напада на поузданост система вештачке интелигенције⁴⁰⁴398

5. Еколошки трошкови коришћења система вештачке интелигенције

³⁹⁶ Urbina F, Lentzos F, Invernizzi C, Ekins S. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3), 189-191. (<https://doi.org/10.1038/s42256-022-00465-9>)

³⁹⁷ Sohn R. 2022. AI Drug Discovery Systems Might Be Repurposed to Make Chemical Weapons, Researchers Warn. *Scientific American* [Internet]. 21 April 2022. See <https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/> (accessed 21 December 2023)

³⁹⁸ Urbina F, Lentzos F, Invernizzi C, Ekins S. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3), 189-191. (<https://doi.org/10.1038/s42256-022-00465-9>)

³⁹⁹ Collins K et al. 2023. ‘Human Uncertainty in Concept-Based AI Systems.’ Paper presented at the *Sixth AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES 2023)*, August 8-10, 2023. Montréal, QC, Canada.

⁴⁰⁰ Su J, Vargas D V, Sakurai K. 2019 One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5), pp.828-841. (DOI: 10.1109/TEVC.2019.2890858)

⁴⁰¹ Verde L, Marulli F, Marrone S. 2021 Exploring the impact of data poisoning attacks on machine learning model reliability. *Procedia Computer Science*, 192, pp.2624-2632. (<https://doi.org/10.1016/j.procs.2021.09.032>)

⁴⁰² Xu Y et al. 2021 Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4). (<https://doi.org/10.1016/j.xinn.2021.100179>)

⁴⁰³ Liu Y, Ning P, Reiter M K. 2011 False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1), pp.1-33. (<https://doi.org/10.1145/1952982.1952995>)

⁴⁰⁴ Verde L, Marulli F, Marrone S. 2021 Exploring the impact of data poisoning attacks on machine learning model reliability. *Procedia Computer Science*, 192, pp.2624-2632. (<https://doi.org/10.1016/j.procs.2021.09.032>)

Прикупљање, анализа, складиштење и дељење података потребних за системе засноване на вештачкој интелигенцији има значајан утицај на животну средину⁴⁰⁵399. На пример, процењује се да складиштење терабајта података годишње потроши 10 kg угљен-диоксида⁴⁰⁶400, док обука за мастер студије права у стилу ChatGPT може створити 550 тона емисије угљен-диоксида⁴⁰⁷401. Процењује се да су глобалне емисије гасова стаклене баште из дата центара исте као и емисије америчке комерцијалне авијације, а како скупови података и модели постају већи, вероватно је да ће се ово повећати. Да би се ублажили негативни утицаји климатских промена, ови системи ће морати да испуне принцип енергетске пропорционалности⁴⁰⁸402 и најбоље праксе еколошки одрживе рачунарске науке (ESCS). Други развоји за побољшање еколошке одрживости алата заснованих на вештачкој интелигенцији укључују:

- Интеграцију техника зеленог рачунарства (green computing) у методе истраживања⁴⁰⁹403
- Стандарди сертификације за одрживе лабораторијске праксе. Ширење мрежа „Зелене лабораторије“ на нивоу базе, у комбинацији са институционалним политикама, може преобликовати норме истраживача⁴¹⁰404.
- Финансијери могу да наложе извештавање о угљенику и подрже надоградње попут енергетски ефикасног хардвера у истраживачким пројектима заснованим на вештачкој интелигенцији⁴¹¹405.
- Влади мандати и регулаторни притисак – правила ЕУ о извештавању о угљенику сада покривају преко 50.000 компанија⁴¹²406.
- Коришћење вештачке интелигенције за оптимизацију и минимизирање утицаја истраживачких метода на животну средину. На пример, имерзивне технологије (виртуелне (VR), проширене (AR) и мешовите стварности (MR)) пружају виртуелна искуства, минимизирајући штету на лицу места и визуелизацију, што може повећати ефикасност припреме^{413,414}407,408

6. Људски трошкови обуке система вештачке интелигенције

Развој и употреба алата вештачке интелигенције ослањају се на критичну, али често невидљиву људску инфраструктуру. Иако је људски рад неопходан за примену вештачке интелигенције, у неким случајевима он остаје недовољно цењен под маском „аутоматизације“. Технолошка критичарка, Астра Тејлор, тврди да се дискурс аутоматизације може користити за маргинализацију одређених доприносилаца научном процесу (нпр. жена или „фантомских

⁴⁰⁵ Henderson P, Hu J, Romof J, Brunskill E, Jurafsky D, Pineau J (2020) Towards the systematic reporting of the energy and carbon footprints of machine learning. *J Mach Learn Res* 21(248):1–43

⁴⁰⁶ Lannelongue L *et al.* 2023 GREENER principles for environmentally sustainable computational science. *Nat Comput Sci* 3, 514–521. (doi.org/10.1038/s43588-023-00461-y)

⁴⁰⁷ Patterson D *et al.* 2021 Carbon emissions and large neural network training. *arXiv preprint*. (doi.org/10.48550/arXiv.2104.10350)

⁴⁰⁸ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023).

⁴⁰⁹ Chithra, J, Vijay, A, & Vieira, D 2014. A study of green computing techniques. *Int. J. Comput. Sci. Inf. Technol.*

⁴¹⁰ Green Your Lab Network. See: <https://network.greenyourlab.org/> (accessed March 21 2024)

⁴¹¹ The Royal Society roundtable on AI and climate science, June 2023.

⁴¹² More than 50,000 companies to report climate impact in EU after pushback fails. *Financial Times*. 18 October 2024. See: <https://www.ft.com/content/a3216188-8e50-4a62-a8d9-e89172b3ddc7> (accessed March 21 2024)

⁴¹³ The Royal Society. 2023 Science in the metaverse: policy implications of immersive technologies. See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/> (accessed 21 December 2023)

⁴¹⁴ The Royal Society interviews with scientists and researchers. 2022 - 2023

радника“) и оправдање мера смањења трошкова без решавања питања равноправности и праведности^{415,416}409,410.

Утицај на рад може се кретати од промена на тржишту рада⁴¹⁷411 до експлоатације радника који раде са подацима и покрећу велике системе вештачке интелигенције⁴¹⁸412. Испитивање организације рада може допринети стварању одговорности за развој одговорних ланаца снабдевања вештачком интелигенцијом⁴¹⁹413.

Решавање питања етике вештачке интелигенције у научним истраживањима

Постоје могућности за научну заједницу (од научника до програмера система и финансијера)⁴²⁰414 да проактивно размотри стратегије за праћење, предвиђање и реаговање на непредвиђене штете узроковане употребом система вештачке интелигенције⁴²¹415.

На основу интервјуа и дискусија за округлим столом, предложене су следеће мере како би се осигурала етичка примена вештачке интелигенције у свим секторима:

- **Таксономија специфична за домен за штете:** Успоставити ревизије, процене утицаја или оквира за евалуацију како би се разумеле социотехничке штете које произилазе из различитих области. Примери укључују таксономију коју је објавио DeepMind, а која наводи различите врсте интеракције човек-рачунар, еколошке и социоекономске штете⁴²²416. Други пример је оквир са више заинтересованих страна развијен за процену друштвеног утицаја генеративне вештачке интелигенције у системима и друштву. Он узима у обзир штете везане за репрезентацију, културне вредности и осетљив садржај, перформансе, приватност и заштиту података, финансијске трошкове, трошкове заштите животне средине и трошкове рада⁴²³417.

- **Етичке смернице и прегледи:** Етичке смернице и кодекси понашања могу водити дизајн модела вештачке интелигенције који се користе за науку, спречити злоупотребу и успоставити најбоље праксе. Примери укључују Хашке етичке смернице⁴²⁴418, које промовишу кодекс понашања ради заштите од злоупотребе хемијских истраживања, или УНЕСКО-ве смернице за етичку вештачку интелигенцију, први глобални стандард о етици вештачке интелигенције, чији је циљ

⁴¹⁵ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

⁴¹⁶ Taylor A. 2018 The Automation Charade. LogicMag. 1 August 2018. See: <https://logicmag.io/05-the-automationcharade/> (accessed February 28 2024)

⁴¹⁷ World Economic Forum, These are the jobs most likely to be lost – and created – because of AI. See: <https://www.weforum.org/agenda/2023/05/jobs-lost-created-ai-gpt/> (accessed February 24 2024)

⁴¹⁸ Barret, M. The dark side of AI: algorithmic bias and global inequality. See: <https://www.jbs.cam.ac.uk/2023/the-darkside-of-ai-algorithmic-bias-and-global-inequality/> (accessed December 10 2023)

⁴¹⁹ Penn J. 2024. Historical review on the role of disruptive technologies in transforming science and society. The Royal Society. See <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>

⁴²⁰ Wang H *et al.* 2023 Scientific discovery in the age of artificial intelligence. *Nature*, **620**. 47-60. (<https://doi.org/10.1038/s41586-023-06221-2>)

⁴²¹ Kazim E, Koshiyama A S. 2021 A high-level overview of AI ethics. *Patterns*, 2(9). (<https://doi.org/10.1016/j.patter.2021.100314>)

⁴²² Weidinger L *et al.* 2022 Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214-229). (<https://doi.org/10.1145/3531146.3533088>)

⁴²³ Solaiman, I *et al.* 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint(arXiv:2306.05949)*

⁴²⁴ The Organisation for the Prohibition of Chemical Weapons (OPCW). The Hague Ethical Guidelines. See: <https://www.opcw.org/hague-ethical-guidelines> (accessed 28 February 2024)

максимизирање користи и минимизирање негативних ризика употребе вештачке интелигенције за научна открића⁴²⁵419.

Потребне су додатне смернице специфичне за поједине области како би се осигурало да научници у различитим областима и секторима могу доносити информисане одлуке приликом интеграције вештачке интелигенције у свој рад.

- **Комуникација и размена знања:** Полазећи од Канцеларије Уједињених нација за питања разоружања⁴²⁶420, компаније из приватног сектора са седиштем у САД (OpenAI, Anthropic, Microsoft, Hugging Face) и цивилно друштво изнели су предлог за побољшање поверења кроз мере изградње поверења, као што су комуникација и координација, посматрање и верификација, сарадња и интеграција и транспарентност⁴²⁷421.

- **Санкције и ограничења:** Истражити регулацију одређеног софтвера и апликација у индустрији и академским круговима, као и одрживост ограничавања приступа алатима и моделима са високим потенцијалом за злоупотребу⁴²⁸422.

- **Укључивање јавности:** Истражити нове приступе управљању како би се укључила погођена јавност у заједничку изградњу ограничења и заштитних ограда. Такође је потребно размотрити стратегију за саопштавање ризика јавности, уз истовремено спречавање општег губитка поверења у науку.

СТУДИЈА СЛУЧАЈА 3

ВЕШТАЧКА ИНТЕЛИГЕНЦИЈА И НАУКЕ О КЛИМИ

Како се могућности вештачке интелигенције и машинског учења даље интегришу у истраживања и примене климатских наука, оне проширују капацитете научника и креатора политике да ублаже климатску кризу⁴²⁹423.

Могућности за вештачку интелигенцију у истраживањима климатских наука

Остваривање потенцијала података климатских наука може бити тешко због хетерогене природе података о животној средини. Ово представља изазов у повезивању података због неусклађених података у просторно-временским резолуцијама; различитих нивоа приватности и безбедности (посебно у вези са сателитским снимањем); и недостатка регулације⁴³⁰424.

Технике учења могу интерполирати празнине у мерењима за сложено препознавање образаца у областима као што је прогноза свемирског времена, техника која се користи у НАСА Центру за

⁴²⁵ UNESCO Recommendation on ethics of artificial intelligence. 2022. See: <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> (accessed 6 February 2024)

⁴²⁶ United Nations, Office of Disarmament Affairs. Confidence Building Measures. See: <https://disarmament.unoda.org/biological-weapons/confidence-building-measures/> (accessed 21 December 2023)

⁴²⁷ Shoker S *et al.* 2023 Confidence-building measures for artificial intelligence: Workshop proceedings. *arXiv preprint(arXiv:2308.00862)*

⁴²⁸ Urbina F, Lentzos F, Invernizzi C, Ekins S. 2023 Preventing AI From Creating Biochemical Threats. *Journal of Chemical Information and Modeling*, 63(3), 691-694 (<https://doi.org/10.1021/acs.jcim.2c01616>)

⁴²⁹ Huntingford C, Jeffers E S, Bonsall M B, Christensen H M, Lees T, Yang H. 2019 Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14(12), 124007. (DOI 10.1088/1748-9326/ab4e55)

⁴³⁰ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>.

климатску симулацију^{431,432}425, 426. Ако се успешно уради, ова фузија скупова података може побољшати тачност модела и процена, доприносећи дугорочном прогнозирању времена, што подржава спремност за катастрофе и управљање ресурсима за екстремне догађаје⁴³³427. Друге технике вештачке интелигенције показале су ефикасност у предвиђању промена средње глобалне температуре⁴³⁴428, предвиђању климатских појава попут Ел Ниња⁴³⁵429, система облака⁴³⁶430 и регионалних временских образаца, као што су падавине у одређеним областима⁴³⁷431. На пример, рад у часопису „Nature“ из 2023. године показао је да је модел вештачке интелигенције предвидео време боље од најнапреднијег система за прогнозирање на свету⁴³⁸432, убрзо након тога, приступ машинског учења компаније DeepMind је надмашио чак и тај критеријум⁴³⁹433. Извештај Краљевског друштва из 2020. године, „Дигитална технологија и планета: Искоришћавање рачунарства за нето нулту потрошњу(Net Zero)“, такође наводи улогу коју вештачка интелигенција може играти у постизању глобалних амбиција нето нулте потрошње⁴⁴⁰434

Вештачка интелигенција за климатску спремност и доношење одлука

Усвајање системског приступа⁴⁴¹435 о вештачкој интелигенцији у климатској науци могло би омогућити ефикасније доношење одлука у вези колиме⁴⁴²436. Климатски модели вођени вештачком интелигенцијом не само да симулирају сложене системе, већ и појачано учење може допринети одрживом креирању политика кроз систематску процену климатских акција,

⁴³¹ Kadow, C, Hall, DM, Ulbrich, U, 2020. Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, **13**, pp.408-413. (<https://doi.org/10.1038/s41561-020-0582-5>)

⁴³² NASA Centre for Climate Simulation. See <https://www.nccs.nasa.gov/news-events/nccs-highlights/acceleratingScience>. (Accessed 21 December 2023)

⁴³³ Buizza, C *et al* 2022. Data learning: Integrating data assimilation and machine learning. *Journal of Computational Science*, **58**, p.101525. (<https://doi.org/10.1016/j.jocs.2021.101525>)

⁴³⁴ Ise, T, Oba, Y. 2019. Forecasting climatic trends using neural networks: an experimental study using global historical data. *Frontiers in Robotics and AI*, **32**. (<https://doi.org/10.3389/frobt.2019.00032>)

⁴³⁵ Ham, YG, Kim, JH, Luo, JJ. 2019. Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568-572. (<https://doi.org/10.1038/s41586-019-1559-7>)

⁴³⁶ Rasp, S, Pritchard, MS., & Gentine, P. 2018. Deep learning to represent sub grid processes in climate models. *Proceedings of the National Academy of Sciences*, **115**, 9684-9689. (<https://doi.org/10.1073/pnas.1810286111>)

⁴³⁷ Zheng, G, Li, X, Zhang, RH, Liu, B 2020. Purely satellite data-driven deep learning forecast of complicated tropical instability waves. *Science advances*, **6**, eaba1482. (DOI: 10.1126/sciadv.aba1482)

⁴³⁸ Bi, K, Xie, L, Zhang, H, Chen, X, Gu, X, Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533-538. (<https://doi.org/10.1038/s41586-023-06185-3>)

⁴³⁹ Wong, C. 2023. DeepMind AI accurately forecasts weather-on a desktop computer. *Nature*. 14 November 2023 (<https://doi.org/10.1038/d41586-023-03552-y>)

⁴⁴⁰ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023).

⁴⁴¹ The Royal Academy of Engineering 2020 Net Zero: A systems perspective on the climate challenge. See raeng.org.uk/publications/reports/net-zero-a-systems-perspective-on-the-climate-chal (accessed 14 October 2020)

⁴⁴² The Royal Society. 2021 Computing for net zero: how digital technology can create a 'control loop for the protection of the planet'. See <https://royalsociety.org/-/media/policy/projects/climate-change-science-solutions/climate-sciencesolutions-computing.pdf> (accessed 21 December 2023)

компромиса и ризика⁴⁴³437. Пример за то укључује дигиталне близанце(*Digital Twin*), виртуелни приказ будућег производа или услуге који се може користити за разумевање, предвиђање и оптимизацију перформанси реалног производа⁴⁴⁴438

Дестинација Земља Европске свемирске агенције развија дигиталног близанца који може генерисати богате токове података како би се омогућила „контролна петља“ за емисије планете⁴⁴⁵439. Ово може олакшати праћење природних и људских активности које доприносе климатским променама, омогућити стручњацима да предвиде и планирају екстремне догађаје и прилагоде политике изазовима везаним за климу⁴⁴⁶440. Примене у стварном свету, попут дигиталног близанца Тувалуа, острвске нације, за заштиту њеног постојања од пораста нивоа мора, показују обећање и виталну потребу за овим технолошким развојем⁴⁴⁷441

Изазови истраживачке етике у климатској науци

Климатска наука заснована на вештачкој интелигенцији има велики потенцијал за решавање изазова глобалних климатских промена. Међутим, овај потенцијал долази са изазовима за истраживачку етику. Идентификација ризика и минимизирање замки остаје кључно за максимизирање позитивног утицаја.

1. Трошкови вештачке интелигенције у погледу животне средине

Спровођење сложених симулација може имати висок угљенични отисак због огромне рачунарске снаге која је потребна, што потенцијално може да поништи очекиване еколошке користи истраживања климатске науке ако се праксе одрживости не интегришу⁴⁴⁸442.

2. Пристрасност и нетачности података

Пристрасност и нетачности у подацима за обуку представљају још један кључни изазов. На пример, модели који су несразмерно обучени на западним популацијама ризикују да превиде проблеме на глобалном Југу и учврсте неправду за регионе у неповољном положају. Континуирано учешће, ревизија и ажурирање система локализованим подацима је кључно за равноправност⁴⁴⁹443

3. Глобално финансирање и грантови

Постоји значајан глобални диспарат у финансирању и расподели грантова. Анализа финансирања истраживања енергије и климе између 1990. и 2020. године показује глобалну неједнакост у расподели финансирања, при чему су западне земље (Европска комисија, Велика Британија и САД)

⁴⁴³ Abrell J, Kosch M, Rausch S (2019) How effective was the UK carbon tax?—A machine learning approach to policy evaluation. SSRN Scholarly Paper ID 3372388. Social Science Research Network, Rochester. 10.2139/ssrn.3372388

⁴⁴⁴ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023)

⁴⁴⁵ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023).

⁴⁴⁶ The European Space Agency. Destination Earth. See https://www.esa.int/Applications/Observing_the_Earth/Destination_Earth. (accessed 21 December 2023)

⁴⁴⁷ Accenture. Case Study: Tuvalu. See <https://www.accenture.com/us-en/case-studies/technology/tuvalu>. (accessed 21 December 2023)

⁴⁴⁸ Henderson P, Hu J, Romof J, Brunskill E, Jurafsky D, Pineau J 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J Mach Learn Res* **21**:1–43

⁴⁴⁹ Royal Society and Department for Science, Innovation and Technology workshop on horizon scanning AI safety risks across scientific disciplines, October 2023. See <https://royalsociety.org/current-topics/ai-data/> (accessed 7 May 2024)

добиле већину средстава⁴⁵⁰444. Приметно је да је у раду утврђено да се ниједна истраживачка институција из Африке није рангирала међу 10 најбоље финансираних институција. Потребна су повећана улагања у иницијативе вештачке интелигенције за недовољно заступљене регионе како би се подржала изградња капацитета и формирање климатских научника на глобалном Југу⁴⁵¹445.

4. Дељење осетљивих података

Подаци о животној средини могу да садрже осетљиве информације, да укључују приватне или личне детаље који се могу повезати са одређеним појединцима или заједницама које нису дале пристанак⁴⁵²446. Просторни подаци високе резолуције и дигитални трагови представљају ризике по приватност, који се увећавају када истраживачима недостаје културно разумевање и осетљивост према различитим заједницама. На пример, иако добронамерна, дигитализација земљишних евиденција ради повећања приступа истраживача подацима може довести до тога да приватни актери попут власника земљишта имају више финансијских средстава да капитализују ове нове податке⁴⁵³447

Еколошки осетљиви подаци такође могу негативно утицати на животну средину⁴⁵⁴448. На пример, дељење података о биодиверзитету, као што су места гнезђења ретких птица, може довести до тога да лоши актери нанесу штету тим срединама⁴⁵⁵449.

Стратегије за етичке истраживачке праксе засноване на вештачкој интелигенцији у климатској науци

1. Тежење пропорционалности енергије

Развити стратегије како би се осигурало да технологије развијене у тежњи ка нето нултој емисији доносе еколошке користи које надмашују њихове емисије⁴⁵⁶450. Интердисциплинарна истраживања о алатима за обрачун угљеника и процену утицаја, попут Пројекта зелених алгоритама⁴⁵⁷451, могу допринети процени и ублажавању утицаја рачунарства на животну средину

2. Побољшање глобалног приступа истраживача подацима

Разлика у приступу истраживача подацима покреће забринутост у вези са правичним развојем и применом вештачке интелигенције⁴⁵⁸452. Ово би могло да омета развој ефикасних климатских решења прилагођених јединственим изазовима одређених заједница. Мреже као што је Одељење

⁴⁵⁰ AbdulRafiu A, Sovacool B K, Daniels C. 2022 The dynamics of global public research funding on climate change, energy, transport, and industrial decarbonisation. *Renewable and Sustainable Energy Reviews*, **162**, 112420. (<https://doi.org/10.1016/j.rser.2022.112420>)

⁴⁵¹ Grantham Research Institute on Climate Change and the Environment. What opportunities and risks does AI present for climate action? See: <https://www.lse.ac.uk/granthaminstitute/explainers/what-opportunities-and-risks-does-ai-present-for-climate-action/>

⁴⁵² Zipper S C *et al.* 2019 Balancing open science and data privacy in the water sciences. *Water Resources Research*, **55**, 5202-5211. (<https://doi.org/10.1029/2019WR025080>)

⁴⁵³ Donovan K P. 2012 Seeing like a slum: Towards open, deliberative development. *Georgetown Journal of International Affairs*

⁴⁵⁴ NBN. Sensitive Data. See <https://nbn.org.uk/sensitive-data/> (accessed 6 March 2024)

⁴⁵⁵ *Ibid.*

⁴⁵⁶ The Royal Society. 2020 Digital technology and the planet: Harnessing computing to achieve net zero. See <https://royalsociety.org/topics-policy/projects/digital-technology-and-the-planet/> (accessed 21 December 2023).

⁴⁵⁷ Green Algorithms Project. See <https://www.green-algorithms.org/> (accessed 21 December 2023)

⁴⁵⁸ National Academy of Sciences. 2024 Toward a New Era of Data Sharing: Summary of the US-UK Scientific Forum on Researcher Access to Data. Washington, DC: The National Academies Press. <https://doi.org/10.17226/27520>.

за статистику за развој Пацифичке заједнице могу промовисати правичан приступ подацима у различитим контекстима, подстичући сарадњу и размену знања⁴⁵⁹453. Слично томе, успостављање поузданих институција за податке може допринети побољшању размене и коришћења података за решавање ванредних ситуација и криза^{460,461}454 455.

3. Контекстуализација управљања подацима

Универзални приступи отвореним подацима не ангажују увек права и интересе мањинских група. Постојећи принципи дељења података попут FAIR (пронађиво, приступачно, интероперабилно, поново употребљиво)⁴⁶²456 могу се допунити принципима управљања усмереним ка људима и сврси, попут CARE принципа за управљање подацима аутохтоног становништва (колективна корист, ауторитет за контролу, одговорност, етика) који разматрају шири приступ осетљивим подацима⁴⁶³457.

⁴⁵⁹ Pacific Community. Statistics for Development Division. See <https://sdd.spc.int/> (accessed 6 March 2024)

⁴⁶⁰ The Royal Society. 2023 Creating resilient and trusted data systems. See <https://royalsociety.org/topics-policy/projects/data-for-emergencies/> (accessed 21 December 2023).

⁴⁶¹ Global Partnership on Artificial Intelligence. 2023 Designing Trustworthy Data Intuitions - Scanning the Local Data Ecosystem in Climate-Induced Migration in Lake Chad Basin - Pilot Study in Cameroon. See: <https://gpai.ai/projects/data-governance/> (accessed 6 March 2024)

⁴⁶² Wilkinson M. D., *et al.* 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9. (doi: 10.1038/sdata.2016.18.)

⁴⁶³ Global Indigenous Data Alliance. Care Principles for Indigenous Data Governance. See <https://www.gida-global.org/care> (accessed 21 December 2023)

ЗАКЉУЧАК

ZAKLJUČAK

Као што је истражено у целом извештају, примене вештачке интелигенције у научним истраживањима доносе ново доба могућности и изазова. Трансформативни потенцијал вештачке интелигенције, подстакнут великим подацима и напредним техникама, нуди значајне могућности у свим областима. Од мапирања дефорестације до помоћи у откривању лекова и предвиђања ретких болести, примене су огромне и обећавајуће. Кроз студије случаја о климатској науци, науци о материјалима и дијагнози ретких болести, овај извештај предвиђа будућност у којој вештачка интелигенција може бити моћно оруђе за научне истраживаче.

Међутим, ове могућности доносе низ изазова везаних за репродуктивност, интердисциплинарни сарадњу и етику. Проналажење равнотеже у којој научници могу искористити предности аутоматизације и убрзани темпо открића, уз истовремено осигуравање интегритета истраживања и одговорне употребе вештачке интелигенције биће од суштинског значаја. Након посвећености Краљевског друштва да се наука – и овом случају вештачка интелигенција – примењује у корист човечанства, извештај позива на колективне напоре у решавању ових изазова. У будућности, и према налазима овог извештаја, три области деловања захтевају пажњу научних заједница и релевантних креатора политике.

Прво је решавање питања приступа и могућности коришћења вештачке интелигенције у науци. Приступ рачунарским ресурсима, висококвалитетним скуповима података, алатима за вештачку интелигенцију и релевантној стручности је кључан за постизање научних продора. У време објављивања, приступ основним инфраструктурама је остао неравномерно распоређен. Ово, заједно са растућим утицајем приватног сектора, као што је истакнуто у поглављу 4., може имати импликације на будућност истраживања вештачке интелигенције на универзитетима. Још један изазов у овој области су силоси знања између стручњака за вештачку интелигенцију и стручњака за научне области (поглавље 3.). Да би се осигурала праведна расподела вештачке интелигенције међу истраживачким заједницама, акције морају ићи даље од олакшавања приступа и фокусирати се на унапређење капацитета за сарадњу, заједнички дизајн и коришћење вештачке интелигенције у различитим научним областима и истраживачким окружењима.

Друго, принципи и праксе отворене науке нуде јасан пут ка побољшању транспарентности, репродуктивности и јавне контроле – све се то показало изазовним у научним пројектима заснованим на вештачкој интелигенцији. Као што је наглашено у поглављу 2., улог нерешавања ових питања је висок, што представља ризик не само за науку већ и за друштво ако примена непоузданих или погрешних резултата заснованих на вештачкој интелигенцији доведе до штете. Потребан је даљи рад како би се разумеле интеракције између отворене науке и вештачке интелигенције за науку и како најбоље минимизирати безбедносне ризике који произилазе из отвореног објављивања модела и података.

Треће, како се улога вештачке интелигенције шири у науци, етичка и безбедносна разматрања морају бити усредсређена у њеном дизајну и имплементацији (поглавље 5). Растуће ослањање на велике скупове података поставља питања о потенцијалној злоупотреби осетљивих информација и предрасудама које би могле да продубе неједнакости или доведу до погрешних закључака. Аутономна природа система вештачке интелигенције такође уводи безбедносне ризике, посебно у областима као што су здравствена заштита или праћење животне средине, где грешке могу имати озбиљне последице; или у областима као што су хемија и биологија, где скупови

података и модели могу пренаменити са злонамерном намером. Решавање ових изазова захтева интердисциплинарну сарадњу и изградњу капацитета научника да предвиде ризик и обезбеде надзор који минимизира потенцијалне штете.

Гледајући унапред, потребна су даља истраживања научне заједнице и креатора политике како би се разумеле импликације вештачке интелигенције на будућност науке. Питања о томе како универзитети могу да прилагоде захтеве за обуком и вештинама, како финансијери могу да наставе да подржавају научни рад који није везан за вештачку интелигенцију и како да оптимизују вештачку интелигенцију за еколошку одрживост кључна су за разумевање утицаја овога на технологију у науци, друштву и на планети.

ДОДАТЦИ

ДОДАТАК 1: Листа слика и боксова које можете погледати у оригиналном тексту

Слика 1. Опис три могућа начина на које вештачка интелигенција може допринети научном разумевању, страна 31 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Слика 2. Трендови подношења патентних захтева за технолошке изуме повезане са вештачком интелигенцијом у последњих 10 година видети слику 2 на страни 67 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf).

Слика 3. Глобална дистрибуција броја патентних породица везаних за вештачку интелигенцију по земљи првог приоритета погледајте на слици 3, страна 68-69 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Слика 4. Глобални тржишни удео машинског учења у природним наукама, по регионима, 2021. (%) погледајте на слици 4, страна 70 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Слика 5. Европски тржишни удео машинског учења у природним наукама, по земљама, 2021. (%) погледајте на слици 5, страна 71 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Слика 6. Репродукција градијента системског приступа коју је развио Hugging Face ⁴⁶⁴350 погледајте на слици 6, страна 77 документа https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai_report_web.pdf

Садржаји BOX-ова су преведени и дата је страница у садржају а овде су BOX-ови у оригиналном тексту:

Box 1 Explainability and interpretability 42

Box 2 Robustness and generalisability in machine learning 47

Box 3 Insights from the Royal Society and Humane Intelligence red-teaming exercise on AI-generated disinformation content 57

⁴⁶⁴ Solaiman, I. 2023 The gradient of generative AI release: Methods and considerations. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 111-122). (<https://doi.org/10.48550/arXiv.2302.04844>)

Box 4 IP Pragmatics 2023: Global patent landscape analysis 66

Box 5 The role of the private sector in patenting medicine and pharmaceutical inventions 74

Box 6 Multilingual language models 83

ДОДАТАК 2

Више детаља о методологији

Резиме истраживачких активности

- Три наручена истраживачка пројекта укључујући историјски преглед улоге деструктивних технологија у трансформацији науке, таксономију употребе вештачке интелигенције у науци, технологији, инжењерству и медицини, и преглед патентног стања вештачке интелигенције и сродних изума.
- Више од 30 полуструктурираних интервјуа
- Четири округла стола на теме репродуктивности, интердисциплинарности, истраживања климатских наука и утицаја великих језичких модела (LLM) у науци.
- Вежба скенирања хоризонта о ризицима вештачке интелигенције за науку, у суорганизацији са Одељењем за науку, иновације и технологију (DSIT)
- Међународни научни форум SAD-UK о приступу истраживача подацима, који су заједнички организовали Краљевско друштво и Национална академија наука

Наручено прикупљање доказа и прегледи

- Penn J, 2024. *Historical review on the role of disruptive technologies in transforming science and society.*
- Berman B, Chubb J, and Williams K, 2024. *The use of artificial intelligence in science, technology, engineering, and medicine.*
- IP Pragmatics, 2024. *Artificial intelligence related inventions.*

Догађаји и истраживачке активности

Краљевско друштво жели да се захвали свима који су допринели развоју овог пројекта, посебно кроз учешће у следећим догађајима.

Више од 30 интервјуа, август 2022 – јун 2023.

Особље Краљевског друштва интервјуисало је научнике и истраживаче из различитих научних дисциплина о новим темама и технологијама у њиховим областима.

Name	Organisation
Aishik Ghosh	University of California Berkley
William Hersh	Oregon Health & Science University
Bruce Weir FRS	University of Washington
Charlotte Deane	University of Oxford
Charlotte Williams FRS	University of Oxford
Conan Donnelly	International Niemann-Pick Disease Registry
David Gao	Nanolayers
Don Canfield ForMemRS	University of Southern Denmark
Doreen Cantrell FRS	University of Dundee
Eileen Furlong FRS	European Molecular Biology Laboratory (EMBL)
Emma Karouna	The Alan Turing Institute
Fran Platt FRS	University of Oxford
Frank Close FRS	University of Oxford
Jamie Rossjohn FRS	Monash University
Jane Francis FRS	British Antarctic Survey
Maria Perez-Ortiz	University College London
Mathieu Denis	International Science Council
Odd Erik Gunderson	Norwegian University of Science and Technology; Aneo
Paul Bates FRS	University of Bristol
Peter Dayan FRS	Max Planck Institute
Peter Hore FRS	University of Oxford
Philip Quinlan	University of Nottingham
Richard Benton FRS	University of Lausanne
Richard Home FRS	British Antarctic Survey
Robin Franklin FRS	Altos Labs Cambridge Institute
Sean Ekins	Collaborations Pharmaceuticals
Simon Boulton FRS	Francis Crick Institute
Stephen Benkovic ForMemRS	Pennsylvania State University
Stephen Smartt FRS	University of Oxford
Tzung-Chien Hsieh	Universitat Bonn

Округли сто о имерзивним технологијама у научним истраживањима, јун 2022.

Краљевско друштво је организовало округли сто на Универзитету у Ексетеру (the University of Exeter), као део серије догађаја „Creating Connections“ који је окупио академике и стручњаке из индустрије са југозападне Енглеске како би разговарали о приоритетима политике за коришћење имерзивних технологија у научним истраживањима. Округлим столом је председавао професор Самјуел Вајн (Samuel Vine), професор психологије на Универзитету у Ексетеру. Кључне теме о којима се разговарало биле су изазови са којима се суочавају истраживачи из индустрије и академици који раде са имерзивним технологијама, укључујући обуку, одрживост и приступ тржиштима. (За више информација погледајте веб страницу Краљевског друштва)⁴⁵⁸.

Name
Gavin Buckingham
Toby de Burgh
Neill Campbell
Kirsten Carter
Danae Stanton Fraser
Stephen Hobbiger
Nathan Mayne
Verity McIntosh
Stephanie Owens
Samuel Vine
Brain Waterfield
Carina Westling

458 The Royal Society. 2023 Science in the metaverse: policy implications of immersive technology. See <https://royalsociety.org/news-resources/publications/2023/science-in-the-metaverse/>

Округли сто о репродуктивности, април 2023.

Округли сто Краљевског друштва о изазовима репродуктивности у научним истраживањима заснованим на вештачкој интелигенцији пружао је увиде професорке Сабине Леонели и Жоел Пино (Sabina Leonelli and Joelle Pineau), као и вишеструку репродуктивност.

Name	Organisation
Dorothy Bishop FRS	University of Oxford
Joaquin Vanschoren	Eindhoven University of Technology; OpenML
Joelle Pineau	McGill University; Meta AI
Malvika Sharan	Alan Turing Institute; Open Life Science
Mark Kelson	University of Exeter
Odd Erik Gunderson	Norwegian University of Science and Technology; Aneo
Ralitsa Madsen	University of Dundee; UK Committee on Research Integrity
Rebecca Kirk	PLOS
Sabina Leonelli	University of Exeter
Sayash Kapoor	Princeton University
Susanna-Assunta Sanson	University of Oxford
Victoria Moody	JISC

Округли сто о вештачкој интелигенцији и климатској науци, јун 2023.

Краљевско друштво је организовало округли сто за климатологе и научнике који се баве подацима како би истражили улогу вештачке интелигенције у истраживањима климатске науке, поделили увиде, изазове и иновативне идеје за будућност ове области. Професорка Џејн Франсис (Dame Professor Jane Francis FRS), чланица Краљевског друштва, пружила је увиде из улоге директорке Британског антарктичког истраживања.

Name	Organisation
Alistair Nolan	Organisation for Economic Co-operation and Development (OECD)
Anil Madhavapeddy	University of Cambridge
Anna Hogg	University of Leeds
Anna-Louise Ellis	Met Office
Dave Topping	University of Manchester
Emily Shuckburgh	University of Cambridge
Jane Francis FRS	British Antarctic Survey
Joycelyn Longdon	University of Cambridge; Climate in Colour
Konstantin Klemmer	Microsoft Research
Philip Stier	University of Oxford
Richard Turner	University of Cambridge
Scott Hosking	British Antarctic Survey; The Alan Turing Institute
Tommaso Venturini	University of Geneva; CNRS
Shakir Mohamed	Google DeepMind
Suman Ravuri	Google DeepMind
Richard Horne	British Antarctic Survey
Timothy Palmer	University of Oxford
Marian Scott	University of Glasgow

Округли сто о интердисциплинарности, јул 2023.

Краљевско друштво је организовало округли сто о улози интердисциплинарности у научним истраживањима вођеним вештачком интелигенцијом. Округли сто је пружио свеобухватно истраживање кључне улоге интердисциплинарности у сналажењу у трансформативном пејзажу научних истраживања вођених вештачком интелигенцијом, са увидима професорке Алисон Нобл (Professor Alison Noble FRS), члана Краљевског друштва, истакнутих стручњака и организација из академске заједнице и приватног сектора.

Name	Organisation
Alison Noble FRS	University of Oxford
Alistair Nolan	Organisation for Economic Co-operation and Development (OECD)
Ankit Agrawal	Northwestern University
Bradley Love	University College London
Cecilia Mascolo	University of Cambridge
Claude Chelala	Queen Mary University of London
Daniele Quercia	King's College London; Nokia Bell Lab Cambridge
Gareth Conduit	University of Cambridge
Georgios Leontidis	The University of Aberdeen
Hujun Yin	University of Manchester
James Dracott	UKRI
Matthias Rillig	Freie Universität Berlin
Michael Castelle	University of Warwick
Mirco Musolesi	University College London
Raffaella Mulas	Vrije Universiteit Amsterdam
Reuben Shipway	University of Plymouth
Seth Baum	Global Catastrophic Risk Institute
Tommaso Venturini	University of Geneva; CNRS
Verena Reiser	Google DeepMind
Victoria Henickx	KU Leuven

Округли сто о великим језичким моделима (LLM)и научним истраживањима, јул 2023.

Краљевско друштво је сазвало округли сто о могућностима и ризицима коришћења великих језичких модела (LLM) у научним истраживањима, на којем су професор Ендрју Блејк (Professor Andrew Blake FRS), члан Краљевског друштва, и Гари Маркус (Gary Marcus) дали уводне речи. Округли сто о коришћењу коришћења великих језичких модела (LLM) у научним истраживањима представио је и позитиван потенцијал и изазове повезане са њиховом интеграцијом. Учесници су нагласили важност развоја стратегија за ублажавање ризика и обезбеђивање уравнотеженог приступа интеграцији LLM-а у истраживање као кључног за одговоран напредак технологија вештачке интелигенције.

Name	Organisation
Alison Noble FRS	University of Oxford
Alistair Nolan	Organisation for Economic Co-operation and Development (OECD)
Andres Guadamuz	University of Sussex
Andrew Blake FRS	Scientific advisor and AI consultant, University of Cambridge
Anthony Cohn	University of Leeds; The Alan Turing Institute
Atoosa Kasirzadeh	University of Edinburgh
Denis Newman-Griffis	University of Sheffield
Edward Tian	GPT Zero
Gabe Gomes	Carnegie Mellon University
Gary Marcus	New York University
Hannah Kirk	University of Oxford; The Alan Turing Institute
Jakob Mökander	Oxford Internet Institute, University of Oxford
James Hetherington	University College London
Jeff Dalton	University of Glasgow
Jessica Montgomery	University of Cambridge
Johan Ordish	Medicines and Healthcare products Regulatory Agency (MHRA)
Matthias Rillig	Freie Universität Berlin
Michael Osborne	University of Oxford; Mind Foundry
Michael Woolridge	University of Oxford
Phil Blunsom	University of Oxford
Samuel Kaski	University of Manchester; Aalto University
Seth Baum	Global Catastrophic Risk Institute
Shreya Rajpal	GuardrailsAI
Yarin Gal	University of Oxford

Радионица о скенирању безбедносних ризика вештачке интелигенције у научним дисциплинама, октобар 2023.

**Уочи Глобалног самита о безбедности вештачке интелигенције, који организује Влада Уједињеног Краљевства, Краљевско друштво ће бити домаћин званичне радионице пре Самита у партнерству са Министарством за науку, иновације и технологију. Догађај је окупио искусне научнике из академске заједнице и индустрије како би скенирали ризике повезане са вештачком интелигенцијом у научним дисциплинама.
(За више информација погледајте веб страницу Краљевског друштва)**

Name	Organisation
Alessandro Abate	University of Oxford
Steven Abel	Durham University
Paul Beasley	Siemens
Craig Butts	University of Bristol
Viscount Camrose	House of Lords, DSIT
Sarah Chan	University of Edinburgh
Linjiang Chen	University of Birmingham
Lee Cronin	University of Glasgow
Gwenetta Curry	University of Edinburgh
Christl Donnelly FRS	Imperial College London
Peter Falkingham	Liverpool John Moores University
Tom Fiddian	Innovate UK
Anthony Finkelstein	City, University of London
Michael Fisher	University of Manchester
Jacques Fleuriot	University of Edinburgh
Ben Glocker	Imperial College London
Julia Gog	University of Cambridge
Seraphina Goldfarb-Tarrant	Cohere
Sabine Hauert	University of Bristol
Cathy Holloway	University College London
Caroline Jay	University of Manchester
Alexander Kasprzyk	University of Nottingham
Frank Kelly FRS	Imperial College London
Rohan Kemp	Department for Science, Innovation and Technology
Georgia Keyworth	Department for Science, Innovation and Technology
Ottoline Leyser	UK Research and Innovation
Richard Mallah	Future of Life Institute
Carsten Maple	University of Warwick
Alexandru Marcoci	Centre for the Study of Existential Risk, University of Cambridge

Continued	
Name	Organisation
Chris Martin	Department for Science, Innovation and Technology
Emran Mian	Department for Science, Innovation and Technology
Daniel Mortlock	Imperial College London
Gina Neff	University of Oxford
Cassidy Nelson	Centre for Long Term Resilience
Thomas Nowotny	University of Sussex
Yannis Pandis	Pfizer
Nathalie Pettorelli	Zoological Society of London
Reza Razavi	King's College London
Yvonne Rogers FRS	University College London
Sophie Rose	Centre for Long Term Resilience
Stuart Russell	UC Berkeley, Future of Life Institute
Abigail Sellen FRS	Microsoft Research Cambridge
Rossi Setchi	Cardiff University
Nigel Shadbolt FRS	University of Oxford
Shaarad Sharma	Government Office for Science
Karen Tingay	Office for Statistics Regulation
Daniel Tor	Department for Science, Innovation and Technology
Mihaela van der Schaar	University of Cambridge
Mark Wilkinson	University of Sheffield

Студија о црвеним тимовима (red teaming-LLMs) мастер студија права за отпорност на научне дезинформације, октобар 2023.

Уочи Глобалног самита о безбедности вештачке интелигенције, који организује влада Уједињеног Краљевства, Краљевско друштво и Хумана интелигенција окупили су 40 постдипломаца из области здравствених и климатских наука како би испитали како потенцијалне рањивости у мастер студијама права ((Meta's Llama 2) могу омогућити генерисање и ширење научних дезинформација (за више информација погледајте веб страницу Краљевског друштва)⁴⁵⁹.

459 The Royal Society. 2024 Red teaming large language models (LLMs) for resilience to scientific disinformation. See <https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/> (accessed 7 May 2024)

Захвалница

Члановима радне групе

Чланови радне групе укључени у овај извештај наведени су у наставку. Чланови су деловали у индивидуалном, а не представничком својству и пријавили су сваки потенцијални сукоб интереса. Чланови су допринели пројекту на основу сопственог стручног знања и добре процене.

Chair

Professor Alison Noble CBE FREng FRS, Foreign Secretary of the Royal Society, and Technikos Professor of Biomedical Engineering, University of Oxford.

Members

Professor Paul Beasley, Head of Research and Development, Siemens.

Professor Peter Dayan FRS, Director, Max Plack Institute for Biological Cybernetics.

Professor Sabina Leonelli, Professor of Philosophy and History of Science, University of Exeter.

Alistair Nolan, Senior Policy Analyst, Organisation for Economic Co-operation and Development.

Dr Philip Quinlan, Director of Health Informatics, University of Nottingham.

Professor Abigail Sellen FRS, Distinguished Scientist and Lab Director, Microsoft Research.

Professor Rossi Setchi, Professor in High Value Manufacturing, Cardiff University.

Dr Kelly Vere, Director of Technical Strategy, University of Nottingham

Royal Society staff

Royal Society secretariat

Denisse Albornoz, Senior Policy Adviser and Project Lead

Eva Blum-Dumontet, Senior Policy Adviser (until July 2023)

Areeq Chowdhury, Head of Policy, Data and Digital Technologies

Nicole Mwananshiku, Policy Adviser

Dr Kyle Bennett, Fast Stream placement

Rebecca Conybeare, UKRI placement

Caroline Gehin, UKRI placement

Рецензенти

Овај извештај су прегледали стручни читаоци и независни панел стручњака, пре него што су га одобрили службеници Краљевског друштва. Чланови панела за рецензију нису били замољени да одобре закључке или препоруке извештаја, већ да делују као независни рецензенти његовог техничког садржаја и презентације. Чланови панела су деловали у личном, а не представничком својству. Краљевско друштво са захвалношћу признаје допринос рецензентима.

Reviewers
Dr Yoshua Bengio FRS, Professor at University of Montreal and Scientific Director of MILA
Ruhi Chitre, Ezra Clark, Tiffany Straza and Ana Persic (Natural Sciences Sector); and Irakli Khodeli (Social and Human Sciences Sector), UNESCO
Dr Rumman Chowdhury, CEO and Founder of Humane Intelligence. 2024 US Science Envoy.
Professor Tony Hey FEng, Honorary Senior Data Scientist at Rutherford Appleton Laboratory. Co-author of <i>Artificial Intelligence For Science: A Deep Learning Revolution</i>



The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- The Fellowship, Foreign Membership and beyond
- Influencing
- Research system and culture
- Science and society
- Corporate and governance

For further information

The Royal Society

6 – 9 Carlton House Terrace

London SW1Y 5AG

T +44 20 7451 2500

W royalsociety.org

Registered Charity No 207043



ISBN: 978-1-78252-712-1

Issued: October 2024 DES8836_5