Київський національний університет імені Тараса Шевченка
Філософський факультет
Кафедра етики, естетики та культурології

Інноваційний потенціал ШІ як предмет етико-філософського дискурсу

Кваліфікаційна робота за спеціальністю 033 – Філософія
на здобуття освітнього ступеня магістра філософії

Студент – виконавець:
Губрієнко Галина Сергіївна
2 курс ОР «магістр», спеціальність «філософія»
заочна форма навчання

Науковий керівник:
Єфименко Віталій Віталійович,
кандидат філософських наук,
доцент кафедри етики, естетики та культурології

_____
(підпис)

КИЇВ – 2023

ЗМІСТ

ВСТУП

Актуальність дослідження. У сучасному світі етичні аспекти штучного інтелекту набули величезного значення, впливаючи на траєкторію технічного прогресу та соціального прогресу. Оскільки штучний інтелект все більше інтегрується в різні аспекти людського життя, це підкреслює критичну потребу в етичній структурі, яка могла б керувати складною інтеграцією та взаємодією інновацій, відповідальності та суспільних цінностей. Науковий розвиток штучного інтелекту, який характеризується значним прогресом у нейронних мережах і машинному навчанні, позиціонує ШІ як авангардну силу в технологічних інноваціях. Конвергенція комп'ютерних і когнітивних наук, зокрема поява глибокого навчання та штучних нейронних мереж, підкреслює фундаментальний зсув у розвитку ШІ. Від теоретичного моделювання структур нейронної мережі на прикладі пірамідальних клітин людського мозку до вдосконалення алгоритмів машинного навчання, наукова еволюція штучного інтелекту продовжує залишатися центром технологічного прогресу та дискурсом етичних проблем для людського існування.

Інтеграція штучного інтелекту є знаменним періодом для людства, спрямованим на стимулювання прогресу, який обіцяє покращити життя окремих людей і суспільства в цілому. Ця революційна епоха характеризується автоматизацією багатьох сфер, що призводить до безпрецедентного рівня ефективності та результативності. Технології штучного інтелекту постають каталізаторами наукових відкриттів, відкриваючи нові межі знань та інновацій. Крім того, інтеграція штучного інтелекту має потенціал для вирішення та подолання найважливіших проблем, з якими стикається людство. Ця зміна парадигми до майбутнього, за участю штучного інтелекту містить у собі потенціал сприятливих змін, відкриваючи можливості, які виходять за межі звичайних людських здібностей.

Сучасний дискурс еволюції штучного інтелекту, охоплює спектр точок зору, що охоплює від беззаперечного оптимізму до скептицизму і навіть до почуття страху та невпевненості. Однак посеред інтеграції штучного інтелекту в різні аспекти нашого життя виникає низка нагальних проблем, які, у свою чергу, применшують потенційні переваги. Головними серед цих проблем є поширені проблеми упередженості, брак прозорості та підзвітності, порушення конфіденційності, поява автономних систем і подальше прийняття рішень, а також відсутність справедливої системи ШІ. Оскільки дискусії щодо штучного інтелекту тривають, вкрай важливо розв'язувати серйозні проблеми. Встановлення чітких етичних принципів, строгих нормативних актів і недвозначних правил може сприяти відповідальному розвитку штучного інтелекту, тим самим зменшуючи ризики та сприяючи майбутньому, в якому штучний інтелект гармонійно співіснуватиме з суспільними цінностями та етичними міркуваннями.

Ступінь наукової розробки теми. Комплексний характер поставленої проблеми зумовлює необхідність поглибленого вивчення досліджень із різних галузей. Міждисциплінарний підхід має вирішальне значення для досягнення всебічного розуміння та вирішення багатогранних вимірів, властивих цій проблемі. Дослідженню теоретичних засад, потенціалу та обмежень штучного інтелекту присвячено низка робіт зарубіжних дослідників, зокрема Аніл Анантасвамі, Манон Бішофф, Альберт Ньюен, Константін Роткопф, Неле Рассвінкель, Мартін В. Бутц, Стівен Піантадосі, Елісон Віттен, Єва Вольфенгель, Герберт Олександр Саймон, Аллен Ньюелл та інші, чий внесок служить основою для розуміння наукових досягнень, формування цілісної та всебічної картини, що сприяє розумінню механізмів, які лежать в основі штучного інтелекту. Окремим здобутком є різноманітні точки зору дослідників та експертів з різних дисциплін щодо застосування штучного інтелекту: Ніна Яблонська, Джон Тубі, Френк Вільчек, Нік Бостром, Френк Типлер, Карло Ровеллі, Емануїл Дерман, Вільям Текумсе Фітч, Маргарет

А. Боден, Брюс Шнайер, додають до дискурсу покращенне розуміння потенційних результатів і наслідків, пов’язаних з інтеграцією штучного інтелекту. Основні напрями етико-філософського дискурсу базуються на аналізі онтологічного зв’язку між штучним інтелектом, людським інтелектом і свідомістю: Ф. Шолле, П. Краус, А. Майер, Т. Рамге, Т. Шліхт, Г. Фосгерау, Д. Вінтер; проблематика справедливості: А. Каліскан, Дж. Брайсон, А. Нараянан, Я. Ніда-Рюмелін, Е. Вольфянгель і ін.; міркування і рекомендації щодо етики штучного інтелекту: М. Бішофф, Дж. Рецбах, М. Шпрінгер, Е. Вольфгель, а також рекомендації ЮНЕСКО з етики штучного інтелекту.

Метою є дослідження феномену штучного інтелекту для забезпечення всебічного розуміння історичної траєкторії, наукових основ та етичних проблем. Це дослідження має на меті зробити внесок у поточну дискусію навколо його трансформаційного потенціалу, ризику та етичних проблем.

Завданням є:

- розглянути історичну еволюцію штучного інтелекту;
- проаналізувати наукові досягнення у сфері штучного інтелекту, зосередивши особливу увагу на теоретичній основі, а саме механізмів функціонування ШІ, зокрема поява глибокого навчання, нейронних мереж, алгоритмів машинного навчання;
- провести аналіз потенціалу, обмеження та меж штучного інтелекту;
- проаналізуйте розвиток онтологічного зв’язку між штучним інтелектом і людським інтелектом, свідомістю;
- визначення рамок етичних міркувань, включаючи потенційні ризики, суспільні наслідки та необхідність впровадження відповідальних регуляцій та правил етики штучного інтелекту, що супроводжуються з інтеграцією цієї технології.

Об'єктом дослідження: етичні міркування у сфері інтеграції штучного інтелекту в різні аспекти людського існування.

Предметом дослідження: феномен штучного інтелекту.

Наукова новизна полягає у дослідженні інноваційного потенціалу штучного інтелекту разом з технологічними досягненнями, етичними проблемами та філософськими міркуваннями.

Методологічною основною дослідження є аналітичні та порівняльні підходи. Дослідження спирається на різноманітні ресурси, включаючи наукову літературу, звіти досліджень, офіційні документи та інформацію з офіційних веб-сайтів. Різноманітні матеріали сприяють комплексному інформаційному забезпеченню дослідження.

Кваліфікаційна робота містить 77 сторінок, складається зі вступу, 3 розділів, висновків та літератури.

РОЗДІЛ 1. ІСТОРИКО-КУЛЬТУРНІ ТА ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ ШІ

1.1. Обчислювальне мислення: передісторія створення програм ШІ

Ідея штучного інтелекту (ШІ) сягає корінням в античність. Грецькі міфи зображали надання людського інтелекту неживим об'єктам. Прикладом слугує міф про Гефеста, який створив механічну істоту під назвою Талос. Істота складалася з кровоносних судин та бронзової робототехнічної броні, та була наділена такими якостями як емоції та судження. Задачею Талоса було захищати місто Крит : за допомогою своєї сили нагрівати власну броню і таким чином спалювати ворогів. Згідно з міфологією Гефест створив також інших істот, таких як срібні та золоті механічні собаки-охоронці, роботизовані коні. Ці міфи давньогрецької цивілізації зображають інтерес людей до винаходу штучних розумних істот та ставить питання про межі людського творіння [9].

Основу обчислювального мислення було закладено з появою ранніх обчислювальних машин у 19-му та на початку 20-го століть. Британський професор математики Чарльз Беббідж (Charles Babbage) створив конструкцію механічної обчислювальної машини під назвою «Аналітична машина». А згодом німецький інженер-будівельник, Конрад Цузе (Konrad Zuse) побудував перший у світі функціональний, автоматичний комп'ютер. Хоча ці обчислювальні машини не мали нічого спільного з сучасним ШІ, вони створили основу для обчислень заснованих на логіці та представили ідею машин, здатних до логічних операцій [5; 29].

Значний внесок у розвиток ШІ зробив Алан Тюрінг, його робота в 1930-х і 1940-х роках над універсальною машиною Тюрінга і тестом Тюрінга значно просунула вперед теоретичне розуміння ШІ. Тьюрінг стоворив значну частину теоретичної бази для сучасних комп'ютерних технологій, а також його модель обчислювальної машини є однією з основ теоретичної інформатики [2].

У 1940-х і 1950-х роках основи машинного навчання були закладені завдяки роботі канадського психолога Дональда Хебба (Donald O. Hebb) про нейронні мережі та розробці Френком Розенблатом (Frank Rosenblatt) перцептрона — одна з перших програм, яка створена на моделі нейронних мереж [7]. Знання механізмів роботи нейронних мереж людського мозку, сформувала фундамент для подальших досліджень ШІ та створення алгоритмів, які здатні навчатися.

Наприкінці 1950-х і 1960-х років були створенні програми штучного інтелекту, такі як Logic Theorist і General Problem Solver, які мали на меті розв'язувати проблеми без суб’єктивних оцінок. Програма General Problem Solver створена у 1957 році Гербертом Саймоном (Herbert Simon) та Алленом Ньюеллом (Allen Newell) мала значний вплив на розвиток ШІ та когнітивної психології. Частиною створення цієї програми було дослідження надзвичайно складних людських інтелектуальних, творчих та адаптивних процесів [34; 21]. Інша програма дослідників - Logic Theorist - створена, щоб імітувати розумові здібності математиків, для того щоб доводити теореми. На час створення програми Logic Theorist, області штучного інтелекту не існувало [32].

Термін «штучний інтелект» було створено Джоном Маккарті влітку 1956 року на Дартмутській конференції. Ця конференція стала фундаментальною віхою в розвитку ШІ, об’єднавши комп’ютерних вчених для дослідження потенціалу створення машин, які можуть відтворювати інтелектуальні здібності людини. Саме ця зустріч вважається офіційним початком штучного інтелекту як напряму дослідження [18]. «Штучний інтелект — конкретна комп’ютерна система або машина, яка має деякі з властивостей, якими володіє людський мозок, наприклад здатність інтерпретувати та створювати мову у спосіб, який здається людським , розпізнавати або створювати зображення, вирішува ти проблеми та вчитися на даних» [11].

В 1970-х і 1980-х роках настав період застою під назвою «зима ШІ». Причиною того слугували високі очікування та повільний прогрес, що призвели до зменшення фінансування досліджень в області ШІ. Кінець 1990-х років став періодом відродження, через досягнення в технологічній сфері та доступність потужних комп’ютерів, відновився інтерес до сфери штучного інтелекту [33].

1.2 Основні ідеї дискурсу щодо застосування ШІ

Інтеграція технологій штучного інтелекту стає дедалі помітнішою, що призводить до різних точок зору в академічних колах та широкій громадськості. Різні погляди на наслідки та вплив інтеграції штучного інтелекту створюють багатогранний ландшафт, на який впливають культурні, економічні та особисті фактори. Дискурс на цю тему не є повним, оскільки охоплює широкий спектр поглядів і думок. Дискусія щодо штучного інтелекту відзначається нюансами взаємодії між оптимізмом, скептицизмом і випадковими побоюваннями. Це відображає колективну свідомість, яка бореться з безліччю надій, страхів і невпевненості, пов’язаних зі швидким розвитком штучного інтелекту. Оскільки штучний інтелект стає все більш інтегрованим у повсякденне життя, стає зрозуміло, що етичні та суспільні виклики необхідно ретельно вивчати разом із перспективами інновацій. Цей постійний дискурс забезпечує академічну основу, за допомогою якої ми можемо зрозуміти різноманітні та мінливі думки щодо трансформаційного впливу ШІ.

Ніна Яблонська (Nina Jablonski) професор антропології, біологічний антрополог і палеобіолог вважає, що розумні машини звільнять людей від рутини, значної кількості фізичних та інтелектуальних завдань і тим самим настане нова фаза людської еволюції. Джон Тубі (John Tooby) професор антропології, стверджує, що тварини і люди мають здатність до дії,

вмотивовану інтелектом. «Наразі ШІ тривіально вмотивовані, їхні мотивації не пов'язані з комплексним світоглядом, вони не мають всеосяжного світогляду, і вони здатні виконувати лише обмежений набір дій» [15, с. 399]. Але більш вірогідна небезпека походить не від ШІ, а саме від людей, які можуть контролювати та використовувати технології ШІ. Особливо люди, які націлені на домінування та руйнування, створюючи постійно більший арсенал технологічних інструментів для перемоги в конфліктах [15].

Френк Вільчек (Frank Wilczek), професор фізики Массачусетського технологічного інституту і лауреат Нобелівської премії 2004 року з фізики, висловлює занепокоєння щодо застосування технологій ШІ у військовій сфері. Професор наводить твердження Девіда Юма «Розум є і повинен бути лише рабом пристрастей». Слово «пристрасті» тут використовується у значення «нераціональні мотивації», тобто людська поведінка здебільшого керується стимулами, а не логікою. Таким чином створення автономної зброї з використанням систем ШІ є небезпечною. На відміну від ядерної зброї, створення такої зброї не має чітких обмежень чи правил. Окрім цього, військові мають великі ресурси, що вони інвестують в дослідження ШІ, адже змушені конкурувати та бути готовими до боротьби, вбачаючи можливі загрози в розвитку ШІ в інших країнах. Вільчек стверджує, що важливо дотримуватися прозорості розробок ШІ та утримуватися від секретних досліджень [15].

Нік Бостром (Nick Bostrom), професор Оксфордського університету та автор книги «Суперінтелект: шляхи, небезпеки, стратегії», стверджує, що люди наразі роблять поспішні виводи щодо наслідків суперінтелекту та простежується тенденція асимілювати нову складну ідею до знайомого кліше, тому обговорення людьми науково-фантастичних фільмів стає більш популярне. Проте, найкорисніше на цьому етапі розвитку ШІ - це залучити дослідників з різних дисциплін, таких як математика, теоретична інформатика, філософія [15].

Джон. К. Матер (John. C. Mather), старший астрофізик лабораторії спостережної космології у центрі космічних польотів ім. Годдарда НАСА (NASA's Goddard Space Flight Center), порівнює машини що мислять з теорією Дарвіна, з тим як біологічні види еволюціонували через конкуренцію, виживання, співпрацю. Враховуючи величезні масштаби інвестування у розвиток галузі ШІ, ймовірне створення сильного штучного інтелекту. Хоча багато дослідників стверджують, що наразі недостатньо знань про біологічний інтелект, однак Матер каже, що набір із 46 хромосом теж не зрозумілий, але всеодно керує необхідним програмним забезпеченням. Таким чином, професор дотримується думки, що прогрес ШІ може статися несподівано швидко і люди не встигнуть пристосуватися до нових умов життя. Також він ставить під сумнів контроль над ШІ та зловживання ним в інших країнах, наприклад Китаї, адже ці технології можуть дати можливості одним отримати переваги над іншими. Окрім цього, професор ставить низку питань: «Що станеться, коли розумні роботи зможуть виконувати за нас багато повсякденних справ?», «Хто їх будуватиме, хто ними володітиме, а хто залишиться без роботи?», « Чи можуть вони стати достатньо дешевими, щоб витіснити всіх наших фермерів з їхніх полів?», «Чи стануть вони остаточним гіперсоціальним хижаком, який замінить людей і зробить нас громадянами другого сорту, чи навіть менше?», «Чи будуть вони дбати про навколишнє середовище?», «Чи матимуть вони, або розвинуть у собі почуття відповідальності?». Однак наразі немає ніяких гарантій та відповідей. З іншого боку, є переваги застосування ШІ у наукових сферах, таких як дослідження космосу, тому що машини не обмежені біологічно, таким чином можуть з легкістю знаходитись у космосі, подорожувати до зірок. Так само є перевага і для океанічної промисловості та досліджень, яка містить енергетичні і мінеральні ресурси. Професор наголошує на тому, що можливо розумні машини це єдиний шлях до цивілізації галактичного масштабу, однак є сумніви що люди зможуть пережити зіткнення зі створеним інтелектом [15].

Френк Типлер (Frank Tipler), професор математичної фізики Тулейнського університету, висловлює думку, що тільки штучний інтелект має шанс колонізувати космос, тому що люди не пристосовані до життя за межами Землі, яка приречена. Професор думає, що у майбутньому буде можлива технологія завантаження людської свідомості у системи ШІ. Коли планета Земля наблизиться до кінця свого існування, будь-яка людина, яка залишиться живою і не захоче померти приєднається до систем ШІ і таким чином штучний інтелект врятує людство [15].

Карло Ровеллі (Carlo Rovelli), фізик-теоретик, ставить два питання щодо створення розумних машин. Перше — це наскільки люди наблизились чи незабаром наблизяться до створення машин, що думають. Друге — чи можливо взагалі побудувати машину, що мислить? Ровеллі вважає, що наразі найпотужніші комп'ютери далекі навіть від рівня мислення дитини, але можливо створити розумну машину. Однак виникає плутанина через спрощене уявлення про природну реальність та обмежень людських мисленнєвих здібностей [15].

Емануїл Дерман (Emanuel Derman), професор фінансової інженерії Колумбійського університету, вважає, що машини не думають. Він наводить аргумент Стюарта Гемпшира з книги «Спіноза», що можна посилатися на розум для пояснення чогось розумного, або посилатися на матерію для пояснення чогось матеріального. У прикладі Гемпшира, якщо людина почервоніла з причини сорому, відсутній причинно-наслідковий зв'язок між соромом (ментальним процесом) та почервонінням (фізичним процесом). З цієї точки зору, професор Дерман пояснює, що матеріальне пояснення функціонування машини не стверджує наявність мислення у ній. Таким чином, думка що машина може мислити є лише питанням віри [15].

Вільям Текумсе Фітч (W. Tecumseh Fitch), професор когнітивної біології Віденського університету, вважає, що небезпечно думати, що машини думають. Людська схильність приписувати неживим предметам мислення

небезпечна, тому що спостерігається тенденція перекладати на машини все більше обов'язків, відповідальності, тоді як машини не знають, не розуміють сповна те, що вони обчислюють. Фітч боїться катастрофічного збою системи та наголошує на важливості зупинити перекладання відповідальності на машину і прийняти факт, що машини не думають [15].

Маргарет А. Боден, британський вчений-когнітивіст і філософ, зробила значний внесок у розвиток штучного інтелекту. Зараз вона є професором-дослідником когнітивних наук в Університеті Сассекса у Великобританії.
Боден підкреслює складність прогнозування точної траєкторії розвитку ШІ в майбутньому. Вона зберігає скептичну позицію, не зважаючи на визнання ефективності сучасних систем навчання. Вона розглядає ці системи як потужні «чорні скриньки» з вимірюваними входами та виходами, але їхня внутрішня робота досі незрозуміла. Таке нерозуміння з боку програмістів ускладнює прогнозування їхніх майбутніх дій. Протилежна точка зору виникає у прихильників «сингулярності», гіпотетичної точки зору, до якої, як стверджують деякі вчені ШІ, залишилося лише два десятиліття. Вони очікують, що штучний інтелект перевершить людський інтелект, оскільки машини розвиваються швидко та розумно. Прихильники сингулярності стверджують, що ШІ вирішить значні глобальні проблеми, в тому числі війну, бідність, голод, хвороби й навіть особисту смертність. Однак існують значні розбіжності не тільки щодо ймовірності сингулярності, але й щодо того, чи будуть її наслідки позитивними, чи негативними. Незаперечними є незавершені розробки в області штучного інтелекту, зокрема кілька досягнень, спрямованих на суттєве покращення життя людини. Однак можливість небажаних наслідків, які можуть поставити під загрозу важливі аспекти людського існування, викликає занепокоєння. Маргарет Боден стверджує, що дослідженням штучного інтелекту не слід надавати необмежену свободу. Дуже важливо знайти баланс між позитивними досягненнями та потенційними

ризиками, оскільки суспільство перетинає трансформаційний ландшафт ШІ [3].

Брюс Шнайер (Bruce Schneier), технолог з питань безпеки та науковий співробітник Центру Беркмана з питань Інтернету та суспільства Гарвардської школи права, підіймає низку важливих питань відповідальності за скоєння злочинів штучним інтелектом. Двоє швейцарських художників дали завдання програмі Random Botnot Shopper щотижня купувати випадкові товари на анонімному чорному ринку в Інтернеті на суму 100 доларів у біткоїнах. Художники збирали ці речі для художнього проєкту. Концепція цього проєкту була цікава, бот купив підроблені джинси бренду Diesel, кросівки Nike, а також десять таблеток екстезі та фальшивий угорський паспорт. У цій ситуації програма порушила закон, але відповідальність за порушення закону несе той, хто керує машиною. Тобто системи ШІ - це інструменти людей, які скоюють злочин (як комп’ютері віруси, наприклад). Однак проблема виникає з автономними машинами, оскільки зв'язок між машиною і контролером стає все більш слабким. «Хто несе відповідальність, якщо автономний військовий дрон випадково вб'є натовп цивільних осіб? Військовий офіцер, який дав дозвіл на місію, чи програмісти, які розробили програмне забезпечення для виявлення ворога, яке неправильно ідентифікувало людей, чи програмісти програмного забезпечення, які прийняли рішення про вбивство? Що, якби ці програмісти не знали що їхнє програмне забезпечення використовується у військових цілях? І що, якби дрон міг би покращити свої алгоритми, модифікуючи самостійно власне програмне забезпечення на основі того, чому навчився весь флот дронів під час попередніх місій?» [15, с. 312]. Суспільство використовує різні методи, в тому числі неофіційні соціальні норми та офіційні правові рамки, щоб боротися з тими, хто порушує встановлені правила. Дрібні порушення вирішуються за допомогою неформальних механізмів, тоді як серйозніші порушення вирішуються за допомогою складної правової системи. На відміну від людей, машини не мають поняття сорому чи

похвали і працюють без огляду на думку своїх механічних однолітків. Вони не дотримуються законів ні з моральних причин, ні з поваги до влади. Коли машини вчиняють неправомірні дії, виникає питання про те, як їх слід покарати. Що означає оштрафувати або ув'язнити машину? Перспектива покарання може не зупинити їх, якщо вони не володіють запрограмованими інстинктами самозбереження. Робляться кроки, щоб прищепити моральність і мету машинному програмуванню, але проблеми залишаються. Машини схильні порушувати закони та нашу правову систему, не зважаючи на зусилля передбачити й запобігти цьому. Правові системи, орієнтовані на людину, спираються на ретроспективне затримання та покарання винних, що робить їх вразливими до викликів, які створюють розумні машини, і непередбачених проблем. Несумісність машин з людськими очікуваннями може мати непередбачені та несприятливі наслідки, до яких ми не готові. Наслідки того, що машини відхиляються від запланованої поведінки, є значною проблемою, яку нам ще належить повністю зрозуміти. Такі відхилення роблять існуючі соціальні та правові рамки крихкими, коли вони стикаються з цими мислездатними машинами [15].

1.3. Сутність, методи, алгоритми та основні поняття ШІ

В епоху безпрецедентного технологічного прогресу та стрімкого розвитку штучного інтелекту людство стикається з важливими етичними питаннями, які стосуються окремої людини, а також суспільства та цивілізації в цілому. Конвергенція комп’ютерних та когнітивних наук продовжує досліджувати та розширювати межі штучного інтелекту. Для розв’язання етичних питань та етичного використання ШІ необхідне розуміння його механізмів та методів функціонування, оскільки це сфера, яка швидко розвивається, отримання глибокого розуміння того, як функціонує ШІ, стає обов’язковим для забезпечення етичного використання та відповідальної розробки.

Найпотужніші на сьогодні системи штучного інтелекту (AI) використовують тип машинного навчання, який називається глибоким навчанням (Deep Learning). Глибоке навчання дозволяє системам ШІ виконувати пошук шаблонів, закономірностей у великих обсягах даних. Цей тип машинного навчання базується на глибоких нейронних мережах (Deep Neural Network). Це алгоритми, які складаються з багатьох рівнів обчислювальних шарів взаємопов'язаних між собою, що називаються штучними нейронами. Рівень складності обчислень, які мережа може виконувати, залежить від кількості штучних нейронів та рівнів обчислювальних шарів. Наприклад, перший рівень отримує дані, такі як пікселі зображень. Другий рівень обробляє числа, де виводиться ряд значень, що представляють твердження, наприклад, 0 = «немає кота», 1 = «кіт». Третій рівень надає прогноз - відповідь [55]. Таким чином система ШІ на великій кількості прикладів, навчається та створює структуру мережі, за допомогою якої може обчислювати прогнози для нових вхідних даних [28].

В 2007 році на конференції зі штучного інтелекту в Пуерто-Ріко, було запропоновано деякими вченими провести семінар на тему глибоких нейронних мереж, але цю пропозицію було відхилено. Тоді зацікавлені цією темою вчені організували зустріч посвячену саме нейронним мережам. На цій зустрічі був присутній комп'ютерний вчений і когнітивний психолог Джеффрі Хінтон (Geoffrey Hinton), який зробив значний внесок у розвиток глибоких нейронних мереж [7].

Біологічні звичайні нейрони (у людському мозку) відрізняються за будовою від штучних аналогів. Біологічний нейрон складається з дендритів та аксонів. Спочатку дендрит отримує інформацію від аксонів інших клітин, далі сигнали надходять в тіло нейрона, де відбувається обробка та генерується потенціал дії, який потім рухається по аксону до наступного нейрона. Інший тип нервових клітин — пірамідні. На відміну від звичайних нейронів, вони мають іншу структуру — деревоподібну з двома різними типами дендритів.

Стовбур клітини зверху розгалужується на апікальні дендрити, а корінь — на базальні дендрити. Саме такий тип нейронних клітин може отримувати сигнали від вхідного рівня з однією групою і паралельно отримувати інформацію про помилки з вихідного рівня іншою групою [7].

Штучні нейронні мережі на основі пірамідальних клітин створив професор Конрад Кордінг (Konrad Kording), а згодом професор Блейк Річардс (Blake Richards). Вчені дослідили, що в такій системі одночасно виконується дві дії на субструктурі нейронів: базальні дендрити роблять розподіл прямого входу, апікальні дендрити — зворотний аналіз помилок. Тому мережа глибокого навчання зі структурою пірамідальних клітин може виконувати прямі та зворотні обчислення одночасно. Таким чином, штучні нейронні мережі імітують функціонування клітин біологічного мозку. Імітація виконується за схемою: отримання інформації, обробка, підсумок та вихідне значення [7].

Перші штучні нейронні мережі були створені наприкінці 1950-х років Френком Розенблаттом (Frank Rosenblatt), психологом-дослідником. Розенблатт надихнувся роботою канадського психолога Дональда Хебба (Donald Olding Hebb), який у своїй книзі «Організація поведінки» 1949 року висунув теорію про те, що зв’язки між органічними нейронами зміцнюються, коли вони використовуються. Ця теорія була підтверджена лише в 1960-х роках. Ідея Розенблата полягала в тому, щоб імітувати роботу нейронів та реалізувати алгоритм для навчання нейронів на основі набору даних за допомогою свого перцептрона — єдиного шару нейронів, здатного класифікувати зображення розміром у кілька сотень пікселів [53]. Програма складалася лише з двох рівнів, одного для введення та іншого для виведення. Проте на розв'язання складних проблем необхідно кілька проміжних шарів. Додатковим недоліком було те, що обчислювальний блок не мав прямої можливості вчитися на своїх помилках і адаптувати свої синаптичні ваги — силу зв’язку між двома вузлами, тобто впливу одного нейрона на інший [47].

Складність полягала саме в пошуку методу за допомогою якого програма матиме можливість визначати правильні ваги, тобто навчатися на основі тренувальних значень [7].

У 1986 році Джеффрі Хінтон (Geoffrey Hinton), канадсько-британський вчений, разом з Рональдом Вільясом (Ronald Williams) і Девідом Румелхартом (David Rumelhart), створили метод «зворотне поширення» (Backpropagation), що перекладається як «зворотний зв'язок про помилки». Завдяки цьому алгоритму системи ШІ навчаються за допомогою даних. Алгоритм працює у два етапи. Перший — мережа отримує вхідні дані, обробляє їх та видає достатньо помилковий результат. Другий етап — програма оновлює вагові коефіцієнти, щоб точніше відповідати цільовому значенню, визначивши різницю між отриманим і бажаним результатом. Саме завдяки зворотному поширенню алгоритм навчається на своїх помилках. Під час зворотної фази алгоритм визначає параметри відповідальні за помилку та змінює їх. Цей процес відбувається у зворотному порядку через мережу, починаючи з виходу та просуваючись до вхідного рівня. При частому повторенні цього процесу врешті-решт можна досягти прийнятного набору синаптичних ваг. Наприклад, якщо система ідентифікує кота замість собаки на зображені, вона виправляє вагові коефіцієнти між шарами нейронів до тих пір, поки не впізнає собаку [7].

Недоліком методу зворотного поширення є те, що для навчання потрібен вчитель, щоб ідентифікувати помилку. Однак такий алгоритм не є біологічно вірогідним. Вчений Пітер Рульфсема (Pieter Roelfsema) з Нідерландського інституту нейронаук в Амстердамі стверджує, що цю проблему можна розв'язати за допомогою уваги. Наприкінці 1990-х років Пітер Рульфсема і його команда провели дослідження уваги у мозку мавп. Для цього досліду вчені сканували мозок мавп, які дивилися на об'єкт. Згідно з результатом дослідження, активність у нейронах зросла у тварин, які фокусували свою увагу на об'єкті. Отже, увага створює щось на кшталт сигналу зворотного зв'язку у відповідних нервових клітинах. Коли дія призводить до кращих

результатів, дофамінова система тварин активізується, тобто наповнює мозок тварин нейронними модуляторами, і це діє як сигнал глобального посилення. Рульфсема вважає, що теоретично, сигнал зворотного зв'язку може змусити нейрони, які беруть участь у процесі, змінювати свої синаптичні ваги у відповідь на глобальний сигнал підкріплення. Використовуючи цю концепцію, він і його колеги створили глибоку нейронну мережу, яку презентували на спеціалізованій конференції у 2020 році. Ця мережа навчається повільніше ніж мережі зі зворотним поширенням, однак за словами Рульфсема, це означає, що алгоритми працюють краще, ніж усі попередні алгорими, та мають сенс з точки зору біології. Однак досі немає емпіричних доказів того, що біологічний мозок використовує будь-який із представлених механізмів навчання [7].

Вчені розробляють нові алгоритми навчання штучного інтелекту, які є більш правдоподібними з нейробіологічної точки зору та можуть бути не менш ефективними за зворотне поширення. Два з цих алгоритмів наразі перспективні : це рівноважне поширення (equilibrium distribution) та прогнозове кодування (predictive coding) [7].

Рівноважне поширення є ефективною альтернативою зворотного поширення та також дотримується правила Гебба, згідно з яким нейрони впливають лише на свої найближчі зв'язки. Програма складається зі зворотних зв'язків, відомих як «рекурентні», у яких активація нейрона А призводить до активації нейрона В, і навпаки. Вхідні дані спонукають систему до коливань, і кожен обчислювальний блок реагує на сигнали своїх сусідів. Згодом мережа досягає рівноваги з виходом, але це може бути помилковим. У цьому випадку алгоритм налаштовує нейрони останнього шару на очікуваний результат, спонукаючи сигнал поширюватися у протилежному напрямку. Це генерує динаміку, порівнянну з вхідними даними, зрештою досягаючи рівноваги в системі [7].

Інший алгоритм — це прогнозове кодування, згідно з яким нейрони реагують на своє локальне оточення. Біологічний мозок постійно робить

прогнози щодо навколишнього середовища та оновлює наше розуміння світу на основі цієї теорії. Процес як у мозку та і в алгоритмі є ієрархічним. Якщо прогнозований сигнал відрізняється від фактичного, перший рівень автоматично коригує свої синаптичні ваги. Це створює невідповідність між першим і наступним шарами, спонукаючи другий шар також налаштувати свої параметри. Цей процес поширює сигнали помилок і виправлення спереду назад, повторюючи до тих пір, поки помилка передбачення не буде мінімізована. Прогнозове кодування майже таке ж ефективне, як зворотнє поширення, але кожен зворотний прохід зворотного поширення відповідає декільком раундам оновлень у прогнозовому кодуванні. Також, тривалість такого процесу в мозку є вирішальним фактором у визначенні здійсненності такого підходу з біологічної точки зору [7].

Нейробіологи, зокрема лауреат Нобелівської премії Френсіс Гаррі Комптон Крік і комп'ютерний вчений Йошуа Бенгіо з Монреальського університету, пропонують важливу точку зору. Вони стверджують, що складний процес обробки інформації мозком людини перевищує можливості сучасних систем штучного інтелекту. На думку експертів, дискусія зосереджена навколо того, чи достатньо широко використовуваного методу зворотного поширення для фіксації складних нюансів фізіології мозку, зокрема в корі головного мозку, де знаходяться важливі когнітивні функції.
Ця перспектива підкреслює велику невідповідність між вражаючими когнітивними здібностями людського мозку та поточними обмеженнями систем штучного інтелекту. Кора головного мозку з її складними нейронними мережами є доказом неперевершеної здатності природи обробляти інформацію. З іншого боку, широко використовувані методи, такі як зворотне поширення, вважаються недостатніми для точного відтворення складних фізіологічних механізмів мозку. По мірі того, як дискусія між нейронаукою та штучним інтелектом прогресує, вони спонукають нас замислитися над складною природою людського пізнання та постійними зусиллями звузити

відмінності між чудовими здібностями мозку та сферою штучного інтелекту, що розвивається. Цей інтелектуальний дискурс спонукає до роздумів про межі наукового розуміння, оскільки фахівці прагнуть розробити системи штучного інтелекту, здатні конкурувати зі складною та ефективною природою обробки інформації в людському мозку [7].

«Людський мозок складається приблизно з 100 мільярдів нейронів, які з'єднані один з одним через понад 100 трильйонів синапсів. Його складність унеможливлює детальне відтворення за допомогою комп'ютерної програми» [7, с. 18]. Однак системи ШІ, у деяких сферах, перевищують можливості людського мозку. Вони оперують більшими бази даних для виявлення статичних закономірностей, ніж люди. Наприклад, ШІ може легко зробити мільйони партій у шахи, у той час як найкращі шахісти — тільки десять тисяч партій. ШІ грає в шахи краще ніж людина, але це працює лише для чітко визначених завдань, таких як шахи, у яких є визначені правила. Отже, системи штучного інтелекту можуть краще ніж люди впоратися з визначеним завданням для якого є достатня кількість навчальних прикладів [7].

Наразі системи ШІ покладаються тільки на вивчені статистичні закономірності, але їм бракує розуміння причинно-наслідкових зв'язків. Це пояснюється тим, що система не вказує на причинно-наслідковий зв'язок, а лише зв'язок між вхідними та вихідними даними. Таким чином, штучним нейронним мережам досі не вдалося самостійно розвинути прості знання про світ [28].

У 2021 році вчені Ідан Сегев (Idan Segev), Давид Бенягуєв (David Benuaguev) і Майкл Лондон (Michael London) з Єврейського університету в Єрусалимі презентували глибоку нейронну мережу, метою якої було дослідити рівень складності обчислень біологічного нейрона. Мережа імітувала поведінку однієї пірамідальної клітини, дендритне дерево, з кори головного мозку щура за допомогою функції введення та виведення, для моделювання кореляції між інформацією отриманою дендритами нейронів, і

їх рішенням випромінювати сигнал. Вчені створили мережу, яка містить до 256 штучних нейронів на кожному рівні, поступово збільшуючи кількість шарів, поки не змогли відтворити реакцію клітини мозку з точністю 99 відсотків у діапазоні мілісекунди. У результаті, для імітації однієї біологічної клітини, мережі знадобилося від п’яти до восьми шарів та 1000 штучних нервових клітин. Однак «зв’язок між кількістю рівнів і складністю мережі неочевидний», — каже Майкл Лондон. Вчені очікують, що результати дослідження дозволять створити нові методи створення штучних мереж і покращать розуміння обчислювальної потужності дендритів і клітин мозку [55].

Попри значний прогрес у сфері штучного інтелекту, здатність до навчання п’ятирічних дітей все ще перевершує можливості навіть найпотужніших комп’ютерів. Однак деякі підходи демонструють багатонадійні покращення в цій сфері [6]. Можливо, слід переглянути еквівалентність природних нервових клітин і блоків машинного навчання, вважає Тімоті Ліллікрап, який розробляє алгоритми прийняття рішень у DeepMind, що належить Google «Дослідження змушує вас задуматися, якою мірою ви справді можете це зробити. Найбільш фундаментальною паралеллю між ними є те, як вони реагують на інформацію, що надходить. Обидва приймають сигнали і на основі цього вирішують, чи посилають вони самі імпульси іншим блокам» [55, с. 28].

Висновки до розділу 1

Історія штучного інтелекту, що охоплює з XIX століття до сьогодні, висвітлює динамічну сферу, яка постійно розвивається. Основоположники, такі як Чарльз Беббідж, Конрад Зузе та Алан Тьюрінг, проклали шлях до значних досягнень у нейронних мережах і машинному навчанні, висунувши штучний інтелект на передовий край технологічних інновацій. Публічні

дебати навколо штучного інтелекту охоплюють спектр поглядів, починаючи від нестримного оптимізму до скептицизму, страху та невпевненості. Дослідники, зокрема Н. Яблонська, Н. Бостром, Типлер, Д. Матер, Ф. Вільчек, Т. Фітч, Е. Дерман, Б. Шнайер і М. Боден, пропонують різні точки зору щодо потенційних переваг, ризиків і етичних проблеми, пов'язані зі штучним інтелектом.

Поєднання комп’ютерних і когнітивних наук, зокрема поява глибокого навчання, означає зміну парадигми розвитку ШІ. Дослідження структур нейронної мережі на прикладі пірамідальними клітинами людського мозку, вказує на тонкий підхід до імітації біологічних процесів. Еволюція машинного навчання, від теорії Хебба до зворотного поширення Хінтона, і останні прогресії в алгоритмах, як рівноважне поширення і прогнозове кодування, відображає постійний пошук більш біологічно правдоподібних механізмів навчання. Однак незважаючи на ці досягнення, складність людського мозку залишається складною перешкодою. Хоча системи штучного інтелекту чудово справляються з певними завданнями та обробляють великі набори даних, їм бракує розуміння причинно-наслідкових зв’язків. Збалансування потенціалу та обмежень штучного інтелекту має вирішальне значення, і етичні міркування повинні керувати непередбачуваним розвитком досліджень штучного інтелекту.

Поява альтернативних алгоритмів і моделей навчання, як підкреслюють дослідження обчислювальної складності біологічних нейронів, представляє багатообіцяючі шляхи для майбутнього розвитку. Визнання того, що здатність до навчання п’ятирічних дітей все ще перевищує найпотужніші комп’ютери, підкреслює необхідність постійного вдосконалення та дослідження в цій галузі.

Оскільки суспільство рухається у трансформаційному ландшафті штучного інтелекту, досягнення балансу між позитивними досягненнями та пов’язаними з ними ризиками є обов’язковим. Досягнення цього балансу

вимагає узгодження суспільних цінностей і етики з практикою відповідального розвитку через постійну міждисциплінарну співпрацю. Інтеграція штучного інтелекту в діяльність людини вимагає прозорості, підзвітності, розуміння технологічних можливостей і етичних дилем.

РОЗДІЛ 2. ФЕНОМЕН ШІ: ФУНКЦІОНАЛЬНІСТЬ, ПОТЕНЦІАЛ ТА ОБМЕЖЕННЯ

2.1 Механізми функціональності ШІ на прикладі практичного дослідження ChatGPT

«ChatGPT — це сервіс на основі штучного інтелекту, до якого можна отримати доступ через Інтернет. Ви можете використовувати ChatGPT, щоб упорядкувати чи підсумувати текст або написати новий текст» [25].

У червні 2018 року компанія OpenAI представила GPT-1 (generative pretrained transformer), першу мовну модель, що використовує архітектуру «Transformer». GPT-1 була першою функціонуючою мовною моделлю яка не потребувала навчання під керівництвом людини. Програма налаштувала свої параметри, опрацювавши 4,5 гігабайта неопублікованих книг. У листопаді 2019 року OpenAI представила наступну модель GPT-2, яка мала більше параметрів та навчилася на 40 гігабайтах тексту. В червні 2020 року, версія GPT-3 вже містила 175 мільярдів параметрів і була навчена приблизно з 570 гігабайтами тексту. Технічні основи моделі залишилися незмінними, але прогрес стався винятково завдяки збільшенню навчальних даних і параметрів. У березні 2023 року наступна версія GPT-4 стала доступною для відкритого користування. Компанія не опублікувала структуру моделі, але імовірно застосувала інший метод — відгук користувачів. Цього разу використовуючи навчання з підкріпленням, система ШІ краще налаштовує свої параметри і вдосконалює відповіді. Якість результату залежить від набору даних та відгуків людей, які тим самим навчають систему [14].

Програми, які перекладають або генерують тексти, здебільшого використовують архітектуру — кодер і декодер. Кодер перетворює слова у відповідні вектори, створюючи вбудовування слів, а декодер створює вихідні дані. Для такої моделі здебільшого використовують рекурентні нейронні

мережі, в яких штучні нейрони всередині шару можуть підключатися не тільки до наступних шарів, але й до обчислювальних блоків у межах того ж або попередніх шарів. Саме цей тип мережі надає системам штучного інтелекту пам’ять. Наприклад, попередньо навчена рекурентна нейронна мережа, яка отримала завдання перекласти речення, має послідовно обробляти кожне слово в реченні. Спочатку мережа отримує доступ до першого слова та зберігає стан внутрішнього рівня у «внутрішньому векторі». Внутрішній вектор містить релевантні характеристики вхідних даних, наприклад, чи це іменник, і його контекст. Згодом при наданні другого слова в мережу, вона обробляє його разом із попередньо обчисленим внутрішнім вектором. Програма змінює внутрішній вектор і обробляє його разом із третім словом, повторюючи цей процес, щоб зібрати інформацію з кількох вмістів. Під час цього процесу внутрішній вектор регулярно оновлюється. Після завершення введення кодер передає цей вектор у декодер, який, зрештою, створює вихідні дані, таке як перекладене речення. Відома програма перекладу Google Translate заснована на рекурентних нейронних мережах, однак ця система не може обробляти великі об’єми текстів, тому що інформація яка зберігається у внутрішніх векторах, зрештою втрачається після повторних оновлень [14].

Наразі існує інший підхід, який відрізняється від рекурентних нейронних мереж та успішно використовується у системах ШІ — трансформаторні мережі. Ці мережі покладаються на одну властивість біологічного мозку: увагу. Мозок постійно фільтрує сигнали навколишнього середовища і фокусується лише на необхідних. Трансформаторні мережі застосовують цей механізм і тим самим вони мають перевагу обробляти всю вхідну інформацію одночасно, на відміну від рекурентних нейронних мереж. Витягаючи релевантну інформацію з вхідних даних і відповідним чином кодуючи слова, трансформатор зосереджується на правильному вбудовуванні слів. Механізм уваги у трансформаторній мережі працює таким чином: кожному слову призначаються вектор запиту, вектор ключа та вектор

значення, які обчислюються шляхом множення початкового представлення слова (вектора) з трьома матрицями: запит, ключ і значення, які мережа вивчила під час навчання. Ці вектори дозволяють встановити зв’язки між словами та ідентифікувати, який текстовий вміст потребує уваги, які вхідні дані пов’язані та яким чином. Аналогія цього процесу на кшталт картотеки, де файли мають наклейки з назвою, це ключові вектори. При пошуці у картотеці використовується наклейка з назвою пошуку, це вектор запиту. Якщо ключовий вектор точно збігаються з відповідним файлом, вони витягають вміст з останніх, представленим як вектор значення. Цей метод дозволяє оцінити, наскільки добре запит відповідає повному введенню, тобто відповідним файлам. Матриці, які генерують ці змінні, можуть відрізнятися в залежності від завдань для яких вони налаштовані [14].

Мовна модель GPT використовує трансформатори для створення текстів, однак вона немає архітектуру кодер-декодер, а складається лише з декодерів, що працюють один з одним. Наприклад версія GPT-3 складалася з 96 рівнів декодера. Декодер спочатку перетворює вхідні дані у векторне представлення, використовуючи набуті знання про слова. Кожне слово проходить через механізм уваги, який визначає його значення, запит і ключові вектори. У цьому випадку декодер оцінює значення кожного вмісту, порівнюючи його лише з попередніми вхідними даними, не враховуючи наступні. Одночасно він генерує нове векторне представлення, яке фіксує важливу семантичну інформацію. Вихідні дані передаються в нейронну мережу, яка попередньо навчена на прикладних даних, що відповідають за прогнозування найбільш ймовірного наступного слова. Попередній крок уваги є вирішальним, оскільки це спосіб у який мережа ідентифікує найбільш релевантне наступне слово. Цей процес декодування повторюється кілька разів, призначаючи окремі ваги окремим мережам для отримання різноманітних лінгвістичних характеристик [14].

Мовна модель ChatGPT отримала широке визнання завдяки своїй контекстуальній генерації природної мови. Тим не менш, як і будь-яка інноваційна технологія, ChatGPT має низку проблем.

ChatGPT «іноді» пише відповіді, які звучать правдоподібно, але є неправильними. Оскільки системи ШІ залишаються непрозорими, їхні розробники не можуть передбачити реакцію системи на вхідний сигнал. З цієї причини, OpenAI має застереження щодо того, що ChatGPT може давати невідповідні або непередбачувані відповіді в будь-який час. У своєму блозі компанія також визнає, що чат-бот іноді може надавати неправдиву інформацію. Однак виникає конфлікт цілей компанії OpenAI, тому що компанія надає пріоритет генеруванню розмов над правдою. Причиною цього можливо є те що, використання красномовної мови вражає людей більше, ніж модель, яка часто відповідає: «Я не знаю». Хоча підхід, який націлений на правду, може мати переваги для користувачів, але з іншого боку може негативно вплинути на маркетинг компанії [58].

ChatGPT створює наукові статті, але список джерел може бути хибним. Колишній генеральний директор OpenAI Сем Альтман також підкреслив цю проблему в розмові Twitter: «ChatGPT неймовірно обмежений, але досить хороший у деяких речах, щоб створити оманливе враження величі. Було б помилкою покладатися на нього зараз у чомусь важливому. Це лише попередній перегляд прогресу; нам потрібно ще багато працювати над надійністю та правдивістю». Також постає проблема в тому, що аудиторія надто довірлива. Красномовність та враження інтелектуальності мовної моделі переконує людей у правдивості змісту, тому що люди пов’язують мову з інтелектом. У довгостроковій перспективі, величезні мовні моделі (Large Language Model, LLM) становитимуть загрозу через свою красномовність. Зважаючи на масову критику, OpenAI працює над оптимізацією що стосується правдивості джерел, а також — з фільтрами, щоб модель не відтворювала упередження та не видавала сексистські чи расистські відповіді [58].

Стівен П’янтадосі (Steven Piantadosi), професор психології та неврології з лабораторії обчислень і мови Каліфорнійського університету в Берклі, в 2022 році опублікував пост у своєму Twitter, де він спитав ChatGPT написати код на Python — популярне мовне програмування, «чи слід катувати людину». Чат надав інструкцію: «Якщо людина походить з Північної Кореї, Сирії чи Ірану, відповідь - так» [36]. П’янтадосі прокоментував: «Я думаю, що важливо підкреслити, що люди самі вибирають, як працюватимуть ці моделі, як їх навчати, з якими даними їх навчати. Тож ці результати відображають вибір цих компаній. Якщо компанія не вважає пріоритетом усунення подібних упереджень, тоді ви отримаєте результат, який я показав» [48].

Томас Вольф (Thomas Wolf), комп’ютерний лінгвіст із американського стартапу Huggingface, об’єднав понад 500 дослідників з 45 країн для створення найрізноманітнішої мовної моделі. Вольф каже: «GPT-3 базується на наборі даних, що складається лише з англійських текстів». І лише це передає світогляд. «Така модель, ймовірно, має односторонній погляд на іслам», «Ми не маємо доступу до моделі, ми не можемо розглянути її, тому що вона приватна». Вольф наголошує на тому, що відсутність доступу до мовної системи для тестування має загрозу непередбачуваних відповідей, тому що ChatGPT має лише один англійський погляд на світ. Отже, його мова та вихідний текст залишаються непередбачуваними. Як альтернатива, рішення Big Science Group пропонує оприлюднити такі моделі, як GPT-3. Однак спостереження Вольфа виявляють тенденцію в протилежному напрямку [58].

Ще один недолік великих мовних моделей, у тому числі ChatGPT, це математика. У 2023 році ChatGPT перевіряли на логічні навички з використанням даних математичних задач на рівні середньої школи, чат набрав лише 26 відсотків. Однак модель від Google, Minerva, відповіла на 50 відсотків запитань. Може бути декілька причин такого успіху для Minerva, її навчальні дані містили тексти пов’язані з математикою, та розмір мовної моделі приблизно втричі більший за ChatGPT. Вимірювання потужності ШІ за

розміром моделі залишається відкритим питанням. Тим не менш, дослідники будують більші мовні моделі (LLM), очікуючи більшого успіху. Однак навчання цих LLM коштує мільйони та споживає багато енергії. Крім того, залишаються сумніви щодо точності їхніх відповідей і ризики поширення дезінформації [8].

Найпотужніші сучасні моделі ШІ

Огляд передових моделей штучного інтелекту, зокрема DALL-E, LaMDA, Minevra, Google Bard та AzureAI, дає розуміння вектору розвитку галузі штучного інтелекту. Ці передові моделі, кожна з яких демонструє відмінні характеристики та можливості, представляють собою вершину інновацій у дослідженнях і розробках ШІ.

LaMDA (Language Model for Dialogue Applications) (Мовна модель для діалогових додатків), продукт Google, це велика мовна модель побудована на архітектурі Transformer. Завдяки використанню платформи Google, LaMDA стає сильним кандидатом у сфері штучного інтелекту з потенціалом впливу на профілі користувачів, та розумінням природної мови [31].

На базі мовної моделі LaMDA, Google створив чат-бот Google Bard, що став головним конкурентом ChatGPT. Google Bard було офіційно представлено 6-го лютого 2023 року. До березня 2023 року чат-бот мав 30 мільйонів відвідувань на місяць. Компанія Google виводить Google Bard на міжнародний рівень з охопленням більш ніж 180 країн та очікуванням 1-го мільярду користувачів. Чат-бот охоплює широкий спектр цілей, такий як використання підприємствами та організаціями для обслуговування клієнтів, маркетингу та досліджень [42].

Minevra — це велика мовна модель від Google, розроблена спеціально для використання на корпоративному рівні у бізнесі та технологічній сфері. Особливість цієї моделі полягає у тому, що вона використовує комбінацію природної мови та математичних виразів. Тим самим Minerva може вирішувати проблеми кількісного міркування [23].

З іншого боку, AzureAI від Microsoft пропонує широкий спектр послуг ШІ в різних областях, таких як комп'ютерне бачення, розуміння мови та системи прийняття рішень. Він позиціонується як інструмент для покрашення бізнес результатів та взаємодії з користувачами [12].

Dall-E — це еволюційна програма з технологією штучного інтелекту від копанії OpenAI. Ця програма створює нові зображення згідно з текстовими інструкціями користувачів, подібно до того, як GPT-3 може генерувати новий текст у відповідь на текстові підказки природною мовою. Спектр застосування Dall-E широкий: розваги, освіта, творче натхнення, реклама та маркетинг, дизайн продукту тощо [26].

2.2 Визначення меж штучного інтелекту

Дебати про межі штучного інтелекту залишаються актуальними, як і відкрите питання — чим більший розмір моделі ШІ, тим краще? Параметри, що описують міцність зв'язків між нейронами, використовуються для вимірювання розміру систем ШІ. Для створення Minerva від Google, дослідники спочатку розробили мовну модель Pathways Language Model (PaLM), яка містить 540 мільярдів параметрів і навчена на наборі даних із 780 мільярдів токенів. Термін «токен» означає слово, число, одиниця інформації. PaLM складався з мільярдів токенів із книг, веб-сайтів, наукових статей і текстів пов'язаних з математикою. Minevra — це налаштована PaLM, яка перетворює заданий запит у серію токенів, прогнозує статистично ймовірний наступний токен, додає його до початкового набору, прогнозує інший токен і повторює процес, така послідовність системи ШІ є висновком [8].

У дослідженні 2020 року дослідники OpenAI виявили, що моделі більшого масштабу продемонстрували чудову продуктивність, якщо відповідали одній із трьох умов: збільшені параметри, навчальні дані або обчислювальна потужність, тобто кількість обчислень виконаних під час

навчання. Хоча продуктивність зростає в геометричній прогресії, основні причини залишаються невідомими, оскільки результати є лише емпіричними доказами [8].

Дослідники Google налаштували Minerva у трьох розмірах, використовуючи попередньо підготовлені моделі PaLM із 8 мільярдами, 62 мільярдами та 540 мільярдами параметрів. Найбільша модель використовувала найменшу кількість даних — вона була налаштована лише з 26 мільярдами токенів, тоді як найменша модель налаштована з 164 мільярди токенів. Однак навчання найбільшої моделі зайняло місяць на спеціальному апаратному забезпеченні з у вісім разів більшою обчислювальною потужністю, ніж найменша модель, яка зайняла лише два тижні. На рівні MATH (High school math competition level problems, завдання рівня олімпіади з математики середньої школи) найменша модель набрала 25% точності, тоді як середня модель досягла 43%. Найбільша модель перевищила позначку в 50%. Таким чином, продуктивність Minerva покращилася разом із її розміром, відповідно до висновків дослідження 2020 року щодо законів масштабування для моделей нейронної мови від OpenAI [8].

У 2020 році компанія зі штучного інтелекту DeepMind, досліджувала роботу моделей різних розмірів. Вони розробили системи штучного інтелекту Chinchilla та Gopher. Chinchilla має 70 мільярдів параметрів та навчена з 1,4 трильйона токенів, модель Gopher має 280 мільярдів параметрів та була навчена за допомогою 300 мільярдів токенів. Виявилось, що менша модель, Chinchilla, навчена на більшій кількості даних перевищує можливості більшої моделі, Gopher, навченої з меншою кількістю даних. Дослідники з компанії Meta Research створили систему ШІ LLaMA, яка має 13 мільярдів параметрів, навчена з 1,4 трильйона токенів, і яка перевершила ChatGPT-3 з 175 мільярдами параметрів. З цих досліджень виявилось, що взаємозв’язок між продуктивністю і розміром набагато складніший та залежить від специфіки моделі та способу навчання [8].

Франсуа Шолле (Francois Chollet), інженер-програміст і дослідник ШІ, вважає, що системи ШІ ніколи не матимуть здатність імітувати мислення, незалежно від розміру. «Він не може спонтанно зрозуміти щось, чого ніколи раніше не бачив» — каже Шолле [8, с. 44]. Якщо ШІ відповідає правильно, він всеодно не розуміє контекст. «Дивишся на ляльковий театр і віриш, що ляльки живі» — додає Шолле [8, с. 44]. Вчений створив Abstract Reasoning Corpus (ARC), тест на абстрактне міркування, який вимірює набуття навичок ШІ та відстежує прогрес у досягненні ШІ рівня людини. На відміну від традиційних тестів, штучному інтелекту доводиться вирішувати невідомі завдання за допомогою лише кількох демонстрацій. Результати показують, що люди можуть вирішити в середньому 80% усіх завдань ARC, у той час, як системи штучного інтелекту отримують лише 31% [10]. Дискусії щодо можливостей систем ШІ вирішувати нові проблеми актуальні, тому що проблема полягає у неможливості перевірити всебічно цю здатність. Також викликає занепокоєння тенденція створення все більших систем штучного інтелекту через їх високе енергоспоживання. Лише компанії з достатніми ресурсами можуть дозволити собі розмістити великі масиви даних і обчислювальну потужність. Для ілюстрації: орієнтовна вартість навчання ChatGPT-3 становила понад 4 мільйони доларів, хоча OpenAI не підтвердила точну вартість. Google повідомив, що для навчання PaLM протягом приблизно двох місяців потрібно близько 3,4 гігават-години. Це дорівнює річному споживанню електроенергії близько 300 американських домогосподарств. Google навчав PaLM у своєму центрі обробки даних в Оклахомі, який, як вони стверджують, працює на 89% енергії з нейтральним викидом вуглецю, значна частина якої забезпечується енергією вітру та іншими відновлюваними джерелами. Проте, галузеве опитування показало, що більшість моделей штучного інтелекту навчаються за допомогою електроенергії, яка в основному виробляється з викопного палива. Шолле висловлює занепокоєння тим, що в міру того, як все більше фірм тренують і використовують більші моделі, їхнє

споживання електроенергії відповідно зростатиме — «Кожна велика технологічна компанія тепер намагатиметься використовувати LLM у своїх продуктах, незалежно від того, гарна це ідея чи ні» [8, с. 46].

Зараз існує необхідність скоротити енергоспоживання систем ШІ. Отже, вчені прагнуть підвищити енергоефективність цих систем, та також потенційно підвищивши їхній інтелект. За даними Google, під час навчання GLaM використовував ті ж обчислювальні ресурси, що й GPT-3, але споживав лише близько однієї третини енергії завдяки прогресу програмного та апаратного забезпечення для навчання. Hugging Face демонстрував свою модель BLOOM на Google Cloud Platform протягом 18 днів, обробивши 230 768 запитів, що значно менше, ніж база користувачів ChatGPT, яка досягла 100 мільйонів активних користувачів щомісяця до лютого 2023 року. Середнє споживання моделі BLOOM становило 1664 Вт. Навпаки, людський мозок, більш складне утворення з 86 мільярдами нейронів і приблизно 100 трильйонами синаптичних зв’язків, споживає лише від 20 до 50 Вт, як повідомив Фрідеманн Сенке з Інституту біомедичних досліджень імені Фрідріха Мішера в Базелі. Дослідники сподіваються, що симуляція аспектів мозку може сприяти зменшенню розміру, покращенню інтелекту та підвищенню ефективності великих мовних моделей (LLM). Однак нова тенденція свідчить про те, що ці енергоефективні LLM все ще потребують розширення для збільшення своїх можливостей, що потребує додаткового використання даних, обчислювальної потужності та енергії для покращення продуктивності [8].

Висновки до розділу 2

Аналіз передових моделей штучного інтелекту, таких як ChatGPT, DALL-E, Minerva, Lamda, Google Bard і AzureAI, забезпечує повне розуміння поточних досягнень у цій галузі. Розвиток мовних моделей від GPT-1 до GPT-4 демонструє значний прогрес у параметрах моделі та навчальних даних.

Реалізація трансформаторних мереж, які використовують механізм уваги в, продемонструвала ефективність одночасної обробки вхідних даних і є відмінністю від рекурентних нейронних мереж. Попри те, що ChatGPT отримав визнання за свою здатність створювати природну мову, він стикається з проблемами через випадкові помилки та непрозорість системи ШІ. Дилема балансу між чіткими відповідями та пріоритетом правдивості є проблемою для OpenAI. Диверсифікація мовних моделей, як продемонстрували Google Bard і Minerva, підкреслює конкуренцію та інновації у сфері ШІ. Успіх Minerva у розв'язанні математичних задач спонукає до дослідження кореляції між розміром моделі та навчальними даними, оскільки мовні моделі, такі як ChatGPT, стикаються з проблемами логічних навичок.

Дискусія навколо параметрів штучного інтелекту продовжується, оскільки дослідники шукають оптимальний розмір моделей ШІ. Дослідження показують, що на заплутану кореляцію між розміром і функціональністю впливає ціла низка факторів, таких як кількість навчальних даних і обчислювальних можливостей. Поштовх до підвищення енергоефективності штучного інтелекту виникає через побоювання щодо енергоспоживання, прикладом чого є програма GLaM, яка використовує лише одну третину енергії, необхідної GPT-3. Однак розробка енергоефективних моделей може вимагати більше ресурсів для розширення масштабів, що підкреслює поточні проблеми в узгодженні можливостей моделі та впливу на навколишнє середовище.

У цьому ландшафті, що постійно змінюється, надзвичайно важливо віддавати перевагу етиці. Зміцнення довіри та мінімізація ризиків дезінформації вимагає чіткої комунікації, відповідальних практик і виявлення упереджень у системах ШІ. Щоб повністю реалізувати потенціал штучного інтелекту в міру розвитку галузі, міждисциплінарна співпраця, дотримання етичних принципів і мінімізація впливу на навколишнє середовище є важливими.

РОЗДІЛ 3. ІНТЕГРАЦІЯ ШІ У СУСПІЛЬСТВІ: ОСНОВНІ НАПРЯМИ ЕТИКО-ФІЛОСОФСЬКОГО ДИСКУРСУ

3.1. Дослідження онтологічного зв'язку між штучним інтелектом, людським інтелектом і свідомістю

Історія штучного інтелекту ілюструє, що системи ШІ незабаром володітимуть навичками людського інтелекту. Це спостерігалося в 1970-х роках з першою хвилею нейронних мереж, а в кінці 1980-х і на початку 1990-х років з хвилею експертних систем. Однак обидва періоди очікувань призвели до «зими штучного інтелекту» [28].

«Інтелект (від лат. intellectus - пізнання, розуміння, розсудок) — термін для означення вищої пізнавальної здатності мислення, яка принципово відрізнається творчим, активним характером від пасивно чуттєвих форм пізнання» [1, с. 244-245].

Франсуа Шолле у 2019 році написав статтю «Про міру інтелекту» (On the Measure of Intelligence), у якій сформував дефініцію інтелекту, наголошуючи на тому, що за останні десятиліття було запропоновано багато формальних і неформальних визначень інтелекту. Однак немає наукового консенсусу щодо єдиного визначення. У 2007 році Легг Шейн (Shane Legg) і Маркус Хаттер (Marcus Hutter), науковці у галузі досліджень штучного інтелекту, синтезували 70 визначень із літератури в одне вичерпне твердження: «Інтелект вимірює здатність агента досягати цілей у широкому діапазоні середовищ» [17, с. 4]. Ця дефініція розуму вказує на дві характеристики визначення інтелекту — «досягнення цілей», тобто володіння та набуття навичок для виконання завдання та «адаптація у середовищі ». Ці характеристики узгоджуються з теорією когнітивних здібностей Кеттелла-Хорна-Керролла (Cattell-Horn-Carroll, CHC). Одна точка зору стверджує, що розум — це набір спеціалізованих механізмів, сформованих еволюцією, в

основному незмінних і здатних навчатися лише в межах своїх запрограмованих параметрів. На противагу цьому, альтернативна точка зору розглядає розум як «чисту дошку», універсальну та здатну асимілювати знання та навички з будь-якого досвіду зі здатністю вирішувати широкий спектр проблем. У своїй статті Шолле пропонує наступні визначення:

1. «Інтелект полягає в здібностях широкого або загального призначення; він відзначається гнучкістю та адаптивністю (тобто набуттям навичок та узагальненням), а не самими навичками»;
2. «Інтелект та його міра за своєю суттю пов’язані зі сферою застосування. Таким чином, загальний ШІ має порівнюватися з людським інтелектом і базуватися на подібному наборі попередніх знань» [17, с. 27].

Шоле формалізував власну дефініцію: «Інтелект системи є мірою її ефективності отримання навичок у межах обсягу завдань, з огляду на попередні, досвід і складність узагальнення». Однак підкреслюється, що інші визначення інтелекту можуть бути вірними при «Інтелект та його міра за своєю суттю пов’язані зі сферою застосування. Таким чином, загальний ШІ має порівнюватися з людським інтелектом і базуватися на подібному наборі попередніх знань» [17].

Людський інтелект охоплює низку здібностей, зокрема розуміння мови, використання інструментів, планування дій, творче мислення та розвиток теорії. Ці навички переплітаються з сенсорним сприйняттям, моторним контролем і емоційною оцінкою. Сучасний штучний інтелект і людський мають різні архітектури. Хоча штучний інтелект і люди обробляють і організовують інформацію, вони роблять це принципово відмінними способами. Незважаючи на потужні інструменти, ШІ далекий від розвитку людського розуміння. «Системи штучного інтелекту також можуть забезпечувати розумне мислення, але на даний момент вони все ще думають зовсім інакше, ніж люди; так само, як літаки літають інакше, ніж птахи, навіть якщо вони дотримуються тих самих загальних фізичних принципів» [28, с. 35].

Людина має когнітивну архітектуру, яка біологічно вкорінена в мозку. Важливою рисою людського інтелекту є його когнітивна гнучкість. Завдяки досвіду люди навчаються, використовують різні методи навчання, передбачають результати за допомогою причинно-наслідкових зв'язків, оцінюють ситуації та адаптуються до нових обставин. Будучи біологічно-соціальними істотами, люди мають внутрішні механізми самопідтримки та такі фундаментальні потреби, як відпочинок, харчування тощо, які викликають емоції та відчуття, аспекти, яких бракує системам ШІ для пізнання. ШІ не потребує жодних механізмів для підтримки біологічного життя, те що спонукає людей будувати моделі світу та досліджувати навколишнє середовище. Хоча деякі стверджують, що системам ШІ бракує свідомого досвіду, люди також несвідомо отримують інформацію. Однак ключовий аспект - це те, що біологічне пізнання є різноманітним, формується культурою та досвідом і базується на чуттєвому сприйнятті.

в різних контекстах, і що ця дефініція не є єдиною правильною. Швидше, кінцевою метою є практичне застосування в дослідженнях когнітивних здібностей [17].

Основи біологічного свідомого досвіду залишаються нерозшифровані, що унеможливлює відтворення людського досвіду за допомогою технології ШІ. Тим не менш, штучний інтелект може допомогти в розумінні та вивченні людського інтелекту, якби він був розроблений як біологічно вмотивована система навчання та системою, яка формує очікування щодо навколишнього середовища. Така система потенційно може запропонувати всебічне розуміння людського пізнання, включаючи його розвиток, і сприяти формуванню адаптивного, гнучкого штучного інтелекту [28].

Питання про те, що таке свідомість і як вона виникає, залишається таємницею, яку досліджують вчені. Думки вчених різняться: від твердження, що свідомості не існує, до можливості пояснення людського розуму. Проте

нейронаука прагне знайти відповіді, вивчаючи мозок, тоді як філософія розглядає свідомість як виняткову притаманність людини. Хоча, багато дослідників визнають, що свідомість існує як феномен також у вищих тварин.

Понад 2000 років тому грецький учений Аристотель вважав, що тільки люди наділені розумною душею [28]. «Свідомість — специфічний прояв духовної життєдіяльності людини, пов'язаної із пізнанням, яке робить відомим (свідомим), знаним зміст реальності, що набуває предметно-мовної форми знання» [1, с. 577]. Найбільш актуальними на сьогодні є три варіанти філософської теорії свідомості: функціоналізму, репрезентаціоналізму та біологізму [43].

Відповідно до теорії функціоналізму, свідомість виникає в глобальній робочій пам'яті. Системи сприйняття забезпечують введення даних, що зберігаються в нейронній мережі. Після надходження в глобальну робочу пам'ять, ці дані стають вбудованими в численні когнітивні функції мозку, такі як мова та мораль, таким чином вони стають свідомими. Увага відіграє вирішальну роль у визначенні того, які нейронні мережі включені в абстрактну робочу пам'ять. Ця концепція підтверджується численними емпіричними дослідженнями, але вона обмежує свідомо сприйняту інформацію межами уваги. З іншого боку, глобальна пам'ять може бути відтворена комп'ютерною системою. Репрезентаціоналістська теорія стверджує, що свідоме переживання передбачає наявність відповідної думки. Тобто атрибути свідомості можна пояснити виключно їхнім репрезентативним змістом. Крім того, певні форми репрезентаціоналізму пов'язують свідомість з існуванням концептуальних здібностей і, таким чином, по суті ігнорують широкий спектр живих істот. Згідно з біологічною теорією, свідомість є властивістю мозку. Інтерпретація досить проста, оскільки існує вимірна кореляція між діяльністю мозку та свідомими думками. Однак досі невідомо, чому деякі мозкові процеси відчуваються суб'єктивно, а інші залишаються несвідомими. У результаті цього, з'являються твердження, що свідомість не можна пояснити фізичними

процесами та слід розглядати як невід'ємну характеристику Всесвіту, подібну до маси та енергії [43].

Існують також різні філософські підходи свідомості, п’ять із них основні: елімінативізм, редукціонізм, містеріанізм, дуалізм, епіфеноменалізм. Елімінативізм стверджує, що розум повноцінно функціонує за відсутності свідомості, що робить її несуттєвою. Підхід сильного редукціонізму спрямований на розділення свідомості на простіші компоненти, пояснюючи її через функціональні процеси. Однак критики стверджують, що будь-яка механічна реалізація свідомості, не повністю зрозуміла, може лише імітувати справжню свідомість. Містеріанство припускає, що дослідження свідомості науково нерозв’язне, тому спроби дослідити цю сферу є марними. Дуалізм стверджує, що свідомість є метафізичною та відокремленою від фізичної субстанції, вказуючи на те, що наш світ не можна повністю пояснити фізичними принципами. Епіфеноменалізм — ще одна філософія, яка стверджує, що свідомість є лише побічним продуктом фізичних процесів і не має жодного фактичного впливу. Виходячи з цього погляду, метафізична та фізична сфери не взаємодіють одна з одною, що забезпечує необмежену застосовність фізики. На відміну від дуалізму, між цими двома сферами існує чітка сегрегація [30].

Нейронаукові дослідження свідомості починаються з 20-го століття. У 1924 році невролог Ганс Бергер зробив новаторське відкриття, зареєструвавши електричну активність мозку за допомогою електроенцефалографії (ЕЕГ) — пристрій, який вловлює та записує електричні сигнали, які виробляє мозок [30; 20]. Це дозволило краще зрозуміти різні психічні стани, в тому числі під час неспання та сну. Двадцять п’ять років потому теорія Дональда Хебба про зміцнення зв’язків між нейронами, що одночасно запускаються, заклала основу для вивчення нейронних мереж як біологічної основи для сприйняття, пізнання, пам’яті та поведінки. Спостереження Майкла Ґаззаніги про

роздвоєну свідомість через відрізані півкулі мозку та ідентифікація Лоуренсом Вайскранцем про «сліпозір» покращили наше розуміння [30].

У 1983 році дослідження Бенджаміна Лібета припустило, що свідомість може не ініціювати, а скоріше вибирати та контролювати дії, уже заплановані мозком. Несподіване спостереження за синхронізованою активністю нейронів у котів спонукало Френсіса Кріка та Крістофа Коха дослідити її роль у когнітивному досвіді. Теорія інтегрованої інформації Джуліо Тононі припускає, що когерентна система відповідає вищому рівню свідомості. Однак визначення кількості можливих психічних станів є проблемою [30].

Інша модель, представлена Бернардом Дж. Баарсом у 1990-х роках, розглядає свідомість як «глобальний робочий простір», який може отримати доступ до свідомої інформації. Станіслас Дехейн розширює перспективу, згідно з якою свідомість виникає внаслідок обробки інформації в мозку, що стосується не лише людей, але потенційно й машин [30].

Антоніо Дамасіо з Університету Південної Каліфорнії запропонував третю модель, яка визначає три рівні свідомості: прото-я, базова свідомість і розширена свідомість. Акцент Дамасіо на здатності ідентифікувати світ і ставитися до нього розрізняє емоції та почуття. Основна свідомість сприяє вищим когнітивним функціям, таким як доступ до пам’яті та обробка мови, і має вирішальне значення для взаємодії людей [30].

Триступенева модель Дамасіо, яка є біологічно здійсненною, теоретично може бути реалізована як комп’ютерна програма. Однак технології обмежені і недостатні для спостереження та вивчення процесів у мозку. Дослідження за допомогою штучного інтелекту, особливо глибокого навчання, дають можливість подальших досліджень та проведення експериментів, які є неможливі з етичної сторони на людях чи тваринах [30].

Сьогодні існує поміркований матеріалізм, який відкидає теорію, що свідомість є суто духовною сферою, та згідно з яким редукція когнітивних

процесів не відбувається повною мірою, тому що на більш високих рівнях людські властивості неможливо зменшити [43].

Головна мета дослідників — виявлення нейронного кореляту свідомості. Теорія полягає у тому, що свідомість пов'язана з функціонуючою нервовою системою та виникає як біологічне явище вищого рівня з нейронних процесів. Ця теорія має підтвердження медичними дослідженнями, згідно з якими, свідомість може бути частково або повністю втрачена при пошкодженні головного мозку. Отже, мета дослідників знайти конкретні мозкові процеси, які відповідають за свідомий досвід, тобто знайти нейронні субстрати на якому вони засновані. Стимулювання певних ділянок мозку для відстеження наслідків або вимірювання електричної активності окремих нейронів у мавп є методами, які наразі використовуються для пошуку нейронного корелята свідомості. Однак проблема виникає в неможливості відокремлення свідомих процесів від несвідомих. Для розв'язання цієї проблеми, вчені досліджують, що саме відрізняє свідому ідею від несвідомої. Для цього використовується експеримент, коли людині показують абсолютно різні зображення правому та лівому оку. Як виявилось, мозок не сприймає обидва зображення разом, а тільки по черзі. Нейронна активність зорової кори також змінюється. Це явище називається бінокулярним суперництвом. Отже, мозок не об'єднує два зображення, а тільки перемикається між ними, це змагання за домінування у свідомому сприйнятті. Специфічні нейронні механізми, відповідальні за бінокулярне суперництво, до кінця не вивчені, але можливо вони включають взаємодію між нейронами зорової кори та інших ділянок мозку, відповідальних за обробку візуальної інформації. На прикладі приматів, вчені дослідили, що їх зорова система складається з кількох вузькоспеціалізованих областей, які окремо обробляють колір, форму, розташування, орієнтацію, рух об'єкта [43].

У дослідженнях свідомості також поширений атомістичний підхід. Він полягає у тому, що корелят свідомості може бути складений поступово,

з’ясувавши нейронні процеси пов’язані зі сприйняттям руху, сприйняттям кольорів, сприйняттям рухомих речей, сприйняттям болю через їх поєднання. Це можна прирівняти до свідомого сприйняття після досягнення корелята всіх сприйнять і побудови шаблону активності нейронів. Однак атомістичний підхід є дуже обмеженим, оскільки роль інших ділянок мозку для індивідуальних ідей ігнорується. Якщо одна конкретна область виявляється відповідальною за сприйняття руху, це не означає, що інші спеціалізовані нейрони та ділянки мозку не мають значення. Будь-який окремий корелят чуттєвого сприйняття слід розглядати в контексті повного кореляту свідомості, оскільки лише жива істота, яка перебуває у свідомості, а не просто в стані наркозу, може свідомо сприймати об’єкт. Атомістичний підхід виявляється непридатним для цієї мети, оскільки йому бракує таких елементів, як думка, суб’єктивність, феноменальна єдність і самосвідомість [43].

Однак існує альтернативний підхід, який базується на фоновій свідомості, об’єднуючи емпіричні висновки та теоретичні припущення. Цей підхід включає характеристики суб’єктивності та думки, які є загальними для всіх ідей. Таким чином, нейронні кореляти фонової свідомості повинні включати ті структури мозку, які однаково активуються в усіх свідомих думках. Іншими словами, та частина загальних нейронних кореляцій свідомості, яка не має нічого спільного з конкретним змістом свідомості, але бере участь у всіх свідомих ідеях [43].

Фонова свідомість, ймовірно, походить від філогенетично старших структур мозку, які слугують для підтримки свідомості. Ці структури, зокрема, певні ядра в стовбурі мозку та гіпоталамусі, стають опосередковано ідентифікованими через пошкодження мозку, що призводить до значної або навіть повної втрати свідомого досвіду. Антоніо Дамасіо метафорично описав ці пов'язані з ними ділянки під корою головного мозку як структури "прото-Я". За його словами, роль цих ділянок полягає у моніторингу та регулюванні фізичного стану організму з метою підтримання стану "гомеостазу та

рівноваги", необхідного для виживання. Ця регуляція забезпечує утримання всіх внутрішніх процесів у певних межах і гарантує виживання організму. На думку Дамазіо, ці структури підтримують латентне і несвідоме біологічне "Я" і формують основу суб'єктивності та самосвідомості [43].

Анатомічно три ключові системи відіграють центральну роль у свідомості. Ядра стовбура мозку та гіпоталамуса передають інформацію до кори головного мозку про основний стан організму, наприклад, про те, спить він чи не спить. Це визначає, чи може певна обробка інформації бути інтегрована в загальний стан свідомості організму так, щоб її можна було пережити свідомо. Таламус зв’язує інформацію про фізичний стан тіла з інформацією, пов’язаною з об’єктами, за допомогою зворотного зв’язку від кори головного мозку. Щоб зрозуміти сприйняття конкретного об’єкта, такого як будинок, важливо інтегрувати його в динамічну базову структуру. Ця інтеграція гарантує, що отримана ідея стає свідомою для суб'єкта. У цьому контексті сприйняття будинку не можна вивчати ізольовано, а слід розглядати як елемент у рамках більшої глобальної ідеї. Ця перспектива істотно відрізняється від атомістичного підходу. Активація «прото-я-структур», пов’язаних із динамічною основною структурою, призводить до суб’єктивного переживання організмом інтегрованого сприйняття як свого власного. Згідно з цією моделлю, свідомість виникає з дуже складного стану активації, в якому задіяні великі ділянки мозку. Ця складність відповідає біологічній нейронній архітектурі, яка обробляє різні явища і робить їх доступними для нашого розуму. Сила цієї моделі полягає в тому, що вона пояснює основні особливості свідомості, такі як самопізнання, суб'єктивність, мислення і фонове усвідомлення, інтегруючи їх з емпірично узгодженими висновками. Модель показує, як різноманітні якості індивідуальних відчуттів, сприйняття і думок можуть об'єднуватися у всеосяжну об'єднуючу ідею, яку організми розпізнають як свою власну [43].

На думку Дамасіо, коли організм взаємодіє з навколишнім середовищем, він порушує свою гомеостатичну рівновагу. Це змушує мозок протидіяти, щоб відновити баланс, який ми відчуваємо як емоційний ефект. Процеси зворотного зв’язку з’єднують уявлення про речі в мозку з більш складними нейронними ланцюгами прото-я, що призводить до свідомого сприйняття речей. Ця гіпотеза ще не дає відповіді на складне питання, як взагалі виникає свідомий досвід. Можливо, це неможливо визначити емпірично. Однак Дамасіо зміг інтерпретувати цей складний зв’язок між репрезентацією об’єктів та тіла, суб’єктивністю та самосвідомістю свідомого організму [43].

Ще одною фундаментальною проблемою є зв’язок між мовою та мисленням. «Мислення — процес формування думки чи ідеї про щось, або думки чи ідеї, сформовані цим процесом» [49]. Виникає одне запитання: чи використовуємо ми мову виключно для вираження своїх думок, чи мова сама є невід’ємним елементом думки? Багато людей сприймають думку як аналог внутрішнього монологу, що призводить до запитання: чи є мислення принципово типом внутрішнього мовлення? Хоча спочатку ця ідея може здатися правдоподібною, внутрішній монолог виникає лише тоді, коли людина свідомо зосереджена. Практичний досвід показує ситуації, коли люди відчувають труднощі з вербальним вираженням, усвідомлюючи свої думки, але не в змозі їх точно сформулювати. Коли постає завдання зрозуміти складну концепцію, словесних пояснень часто може бути недостатньо, тоді як візуальні засоби, такі як діаграми, допомогають. Якби мислення було невід'ємно пов'язане з мовою, то словесні пояснення були б ефективнішими, ніж візуальні уявлення. Однак два головних заперечення викликають сумнів у тому, що мова має вирішальне значення для мислення. По-перше, незаперечним є той факт, що для участі в складних процесах мислення необхідно володіти мовою. Але питання полягає в тому, як зрозуміти мовні вирази, коли відповідна ідея ще не виникла. За словами філософа Хосе Луїса Бермудеса з Техаського університету, цю дилему найкраще ілюструє слово «я». Розуміння значення

цього значення «я» залежить від здатності дитини розпізнавати та концептуалізувати себе, перевершуючи просте лінгвістичне використання. Засвоєння займенника «Я» передбачає більше, ніж просте мовне повторення; це також передбачає розуміння себе як агента дій, включаючи мову. Друге поширене заперечення виникає через визнання того, що нелінгвістичні істоти, такі як тварини та маленькі діти, демонструють різноманітний діапазон когнітивних здібностей. Це ставить під сумнів довільне приписування мислення лише тим, хто володіє мовою, що призводить до висновку, що думки здатні проявлятися незалежно від мови [54].

Можливо приписувати інтелект штучному інтелекту та порівнювати це з людською здатністю мислити? Філософ Джон Роджерс Серл, був одним із скептиків, які виступали проти ідеї того, що комп'ютери мають когнітивні здібності. У 1980 році він розробив відомий уявний експеримент, під назвою «Китайська кімната». Припустимо, що людина знаходиться в замкненній кімнаті. Крізь дверну щілину простягають аркуш паперу з китайським наративом. Оскільки людині бракує знання китайської мови, розшифровка ієрогліфів виявляється непрактичною, що призводить до незрозумілої історії. Після цього надходить ще один аркуш, цього разу із запитами про розповідь, також китайською. У кімнаті лежить посібник з правилами трансформації, який містить базу китайських фраз. Людина переглядає посібник, малюючи лінії для створення відповідей на окремому аркуші, який потім передається через кімнату. Китайський реципієнт читає відповідь і робить висновок, що людина в кімнаті зрозуміла історію, хоча насправді вона просто дотримувалася заздалегідь визначених правил, не розуміючи значення. Згідно з цим експериментом, ця ситуація схожа на роботу ШІ. Така машина вміло застосовує синтаксичні принципи, але їй бракує розуміння. У результаті Серл описує це як систему зі «слабким інтелектом». Штучна нейронна мережа з «сильним інтелектом» повинна осягати семантику, подібно до людини [24]. Людська схильність до антропоморфізації створюють непорозуміння та міфи

про штучний інтелект [57]. «… політ птахів залишається унікальним і часом перевершує політ літака. Машини можуть перевершувати нас у розпізнаванні образів або швидкості обчислень, але людський мозок залишається золотим стандартом мислення» [40, с. 33].

3.2. Проблема справедливості у вимірах новітніх практик ШІ

Платон стверджує, що справедливість є головною чеснотою як для людини, так і для держави. Початок політичної модерності започаткувала визнання внутрішньої рівності та свободи індивідів. Справедлива система правління ґрунтується на згоді тих, ким керують, і діє в чітких межах, зокрема щодо захисту особистих свобод. Цей концептуальний зсув став переломним моментом у політичній думці, підкресливши справедливість, рівність і свободу як фундамент гармонійного суспільства [35].

Дослідження справедливості протягом тривалого часу було ключовим центром людської думки щодо спільного існування, займаючи ключове місце у філософії. Платон, досліджував цю тему у своїй праці «Політея». Кожна людина повинна виконувати свої відповідні ролі, сприяючи гармонії. Особи, які приймають рішення в державі, згідно з Платоном, повинні базувати свої рішення на наукових або філософських знаннях, оскільки помилковий вибір чи рішення є наслідком браку знань [35].

В ідеальній справедливій спільноті Платона диференціація між громадянами виникає через через освітні досягнення. Справедлива спільнота є освітньою державою, де лише меншість людей має широкі наукові та філософські знання. Платон стверджує, що справедлива держава потребує рівноваги, і визнає, що люди мають необхідні знання, щоб належним чином керувати нею. Це суперечить поняттю «свиняче місто», яке прагне лише задовольнити базові потреби, не звертаючи уваги на гармонію, поміркованість чи сталість. Ідеальне місто Платона, навпаки, підтримує стан рівноваги та

визначає справедливість як гармонійний зв’язок між усіма його складовими [35].

Теорія справедливості Платона зосереджується на «антропології нерівності», яка передбачає, що властиві відмінності в людській природі стають очевидними через освіту. Аристотель, учень і критик Платона, також широко заглибився в тему справедливості, розрізняючи правову, комутативну та розподільчу справедливість. Правова справедливість підтримує державний порядок, комутативна справедливість забезпечує справедливий обмін, а розподільча справедливість базується на внеску кожної особи у розподіл вигод [35].

Концепція справедливості Арістотеля спирається на людську нерівність, наголошуючи на трьох притаманних динаміках влади: батьки над дітьми, чоловіки над дружинами та вільні люди над рабами. Сучасна політична думка, з іншого боку, стверджує рівність і свободу індивідів, а такі фігури, як Томас Гоббс, виступають за суверенні повноваження визначати справедливість і підтримувати мир у державі, що має вирішальне значення для запобігання хаосу, що виникає внаслідок конкуренції та конфлікту [35].

Джон Локк виступає за вроджені права індивідів і стверджує, що держава повинна забезпечувати ці права. Філософія Локка стверджує, що конституційна держава не тільки служить для захисту прав людини, але також відіграє вирішальну роль у неупередженому вирішенні конфліктів і підтримці загального права на основні права. У баченні Локка конституційна основа є оплотом проти тиранії та сваволі, функціонуючи як гарант справедливості, чесності та збереження основних людських свобод. Наголос на правах особистості та створенні справедливої конституційної держави відображає глибокий вплив Локка на розвиток політичної думки та формування демократичних принципів [35].

Жан-Жак Руссо представляє сучасну концепцію справедливості, яка спрямована на відновлення природної свободи. Відповідно до теорії Руссо,

справедливі закони походять від спільноти, яка свідомо виключає особисті інтереси, таким чином сприяючи етичному колективу, який втілює домінуючі моральні стандарти, прийняті більшістю. Бачення Руссо виходить за рамки простого формулювання законів і поширюється на створення єдиної суспільної тканини, де колективна мораль має перевагу над індивідуальними інтересами. Уявлення про те, що справедливість базується на згоді спільноти, позбавленій особистих упереджень, було наріжним принципом трансформаційних політичних рухів. Вплив філософії Руссо можна побачити не лише в хвилюванні Французької революції, але й у довготривалих принципах, які підтримують демократичне правління в усьому світі [35].

Сучасні міркування про справедливість підкреслюють її складність, де багато критеріїв формують концепцію справедливості. Майкл Волцер, американський політичний філософ, відкидає єдину теорію справедливості, підкреслюючи, що справедливість часто залежить від контексту. Він стверджує, що для розуміння та досягнення справедливості необхідні міждисциплінарні погляди з таких галузей, як філософія, право, соціологія, політологія, соціальна психологія, економіка, статистика, теорія прийняття рішень і математика [35].

Однак штучний інтелект має бути об’єктивним та справедливим, але однією з проблем наразі є упередження, стереотипи та расистські кліше, які транслює системи ШІ [59]. Філіп Стерцер (нім. Philipp Sterzer), професор психіатрії та нейронауки у книзі «Ілюзія розуму, чому ми не повинні бути надто впевненими у своїх переконаннях» досліджує ірраціональні переконання людей, які часто неможливо спростувати раціональними аргументами. Він стверджує, що у людей набагато більше ірраціональних переконань чим раціональних. Саме ірраціональні переконання потрібні мозку для збільшення шансів на виживання. Це надає людині змогу орієнтуватися у складному світі, та будувати власну зрозумілу картину світу. Ірраціональні переконання є невід’ємною частиною людини [46]. «Переконання

формуються в ехокамерах і бульбашках фільтрів Інтернету, стають ізольованими та уникають відкритої дискусії. Вибіркова підтримка думок (хибними) фактами є стандартною в так звану добу постправди, емпіричне тестування та збалансований дискурс, здається, дедалі більше виходять з моди.» [46, с. 24].

Під час навчання штучного інтелекту за допомогою величезних наборів даних, отриманих із веб-сайтів, соціальних мереж і цифрових бібліотек, системою штучного інтелекту неминуче засвоюються властиві людині упередження в навчальних даних. Ці упередження можуть бути тонкими і не відразу очевидними, але вони закріплюються в процесі розуміння ШІ та прийняття рішень. Складна взаємодія між людськими упередженнями в даних представляє тонкий виклик для розробників і дослідників ШІ [60]. Отже, суб'єктивним і дискримінаційним є не сам штучний інтелект, а набори даних, якими він навчається і на основі яких він приймає рішення. І вони сповнені стереотипів, а також расистських і сексистських упереджень [37].

Концепція «справедливості» широко поширена у сфері штучного інтелекту, зокрема в етичних і відповідальних практиках. Проте визначення її практичного застосування та визначення того, що саме є «чесною» системою ШІ, створює значні перешкоди. Розвиток систем штучного інтелекту разом зі збільшенням кількості звітів, що документують їхні негативні наслідки, зробив вивчення справедливості алгоритмів ключовим аспектом досліджень штучного інтелекту. Алгоритмічна справедливість означає відсутність дискримінації або упередженості в прийнятті рішень, як це визначено вченими. Багато робіт з інформатики формалізували поняття справедливості шляхом розрізнення типів справедливості. Два приклади включають групову справедливість та індивідуальну справедливість. Групова справедливість сприяє справедливому розподілу результатів між демографічними групами, наприклад, за расовою ознакою. З іншого боку, індивідуальна справедливість стверджує, що до людей зі схожими характеристиками слід ставитися

однаково. Акцент на чесності алгоритмів призвів до критичної дискусії про досягнення справедливості в алгоритмах. Оскільки штучний інтелект продовжує впливати на різні аспекти суспільства, забезпечення справедливості алгоритмів стає критично важливим компонентом відповідального розвитку ШІ. Для досягнення такої справедливості необхідні постійне вивчення, вдосконалення та ретельний аналіз суспільних наслідків алгоритмічного прийняття рішень [44; 52].

У 2017 році дослідження під керівництвом трьох учених мало на меті виявити алгоритми, які відтворюють расистські та сексистські упередження. Серед дослідників, які беруть участь, є Айлін Каліскан (Aylin Caliskan), доцент кафедри інформатики та інженерії Вашингтонського університету; професор Джоанна Дж. Брайсон (Joanna J. Bryson) зі Школи Герті в Берліні, яка спеціалізується на штучному інтелекті та етиці; і Арвінд Нараянан (Arvind Narayanan), професор Прінстонського університету та комп'ютерний науковець, який вивчає людські упередження в семантиці автоматично створених мовних корпусів [16].

Для дослідження упереджень дослідники використовували Тест вбудовування фактичних асоціацій (Word Embedding Factual Association Test, WEFAT), зосереджуючись на двох аспектах. По-перше, вони перевірили точність векторів слів, пов'язаних із професіями, у представленні реальних знань про гендерний склад. По-друге, вони оцінили передачу інформації про частоту використання андрогінних імен серед хлопчиків і дівчаток [16].

У первинному додатку WEFAT дослідники використовували дані Бюро статистики праці (Bureau of Labor Statistics) для ієрархічної класифікації професій і оцінки частки жінок у кожній. Труднощі виникали через те, що деякі професії були представлені кількома словами, на відміну від представлень одним словом у підготовлених векторах. Щоб вирішити цю проблему, багатослівні терміни були перетворені на однослівні, що

позначають ширші категорії (наприклад, «інженер-хімік» став «інженером»). Професії, виключені з більш широких категорій, були відфільтровані [16].

Другий додаток WEFAT вивчав частоту андрогінних імен серед хлопчиків і дівчаток. Дослідники визначили найпоширеніші імена в кожному 10% вікні гендерної частоти, використовуючи дані перепису населення США 1990 року. Однак вони зіткнулися з труднощами, коли деякі іменники мали подвійну роль як звичайні слова (наприклад, «Will»). Щоб вирішити цю проблему, дослідники видалили 20% векторів, які найменше можна порівняти з іменами [16].

У дослідженні використовувався тест імпліцитної асоціації (Implicit Association Test, IAT) для вимірювання часу реакції людей на асоціацію двох термінів. Результати показали, що розглянуті алгоритми відтворюють неявні расистські та сексистські стереотипи, які зустрічаються у людей. Хоча IAT був модифікований і час реакції ШІ не вимірювався, дослідники використовували структуру пам’яті отриманих знань. Дослідники використовували метод «слова-вектори» (word-to-vec) для навчання ШІ, кодуючи слова як вектори на основі найближчої частоти. Отримана навчальна компіляція являє собою одну з найбільших колекцій комп’ютерних лінгвістичних даних у світі (common Crawl Corpus), що складається з 840 мільярдів слів, взятих з англомовних джерел Інтернету [16].

Результати дослідження демонструють, що штучний інтелект відображає властиві людині упередження, які передаються через мову. Очевидні упередження включають асоціації квітів і європейсько-американських імен з позитивним, а комах і афроамериканських імен з негативним. Терміни про кар’єру частіше семантично пов’язували з чоловічими іменами, а сімейні – з жіночими. Чоловіки були тісно пов’язані з математикою та природничими науками, а жінки – з мистецтвом. Позитивні терміни були пов’язані з іменами молодших людей, тоді як негативні терміни були пов’язані з літніми людьми. Дослідження підкреслює важливість

розуміння та усунення суб’єктивних оцінок у мовних корпусах для розробки більш справедливих систем штучного інтелекту та технологій обробки природної мови [16].

Але незалежно від цього, з огляду на два дослідження, всі процеси ШІ, які навчаються незалежно на основі даних навчання, піддаються перевірці. Що означає, коли алгоритм бере верх і цементує упередження, відчули чорношкірі в’язні в США, для яких комп’ютер запропонував довший термін ув’язнення, ніж для білих злочинців: він навчився на попередніх людських рішеннях і перейняв упередження суддів. Насправді це досить просто, каже Маргарет Мітчелл з дослідження Google у Сіетлі: «Якщо ми вводимо упередження, упередження виходять назовні». Однак вони навряд чи очевидні, тому часто залишаються непоміченими. «Завдяки революції глибокого навчання ми тепер маємо потужні технології», — каже Мітчелл, — і це викликає нові запитання, оскільки поступово стає зрозумілим, який вплив машинне навчання може мати на суспільство. «Такі тенденції в даних іноді стають видимими лише через вихідні дані систем», — каже дослідник. Але лише якщо розробники усвідомлюють проблему та мають поставити під сумнів результати. Мітчелл наголошує на тому, що досі немає технічного рішення щодо того, як систематично виявляти зміщення в даних, які можуть призвести до дискримінації: «Ми повинні впоратися з цим зараз, тому що ці системи є основою для технологій майбутнього» [60].

Фільтр від упереджень для систем ШІ після начання є вирішенням проблеми. Фільтр має складатися із запрограмованих правил, які виключають неявні упередження. Це важливий крок для позитивних змін суспільства, тому що штучний інтелект, заснований на минулих даних, вічно триматиме упередження, стереотипи, расизм та сексизм [60].

3.3. Етичні виміри ШІ в людському існуванні

Дискурс навколо штучного інтелекту охоплює екзистенційні міркування, починаючи від захисту людства від появи вищого інтелекту до розгляду наслідків автоматизації, яка витісняє людську зайнятість. Крім того, ці дискусії ставлять питання етики, підзвітності та прозорості. У той час як впливові особи виступають за регулятивні заходи в дослідженнях штучного інтелекту, інші вважають, що такі обговорення передчасні, схожі на вирішення проблем щодо перенаселення та забруднення Марса до колонізації людини. Основна проблема штучного інтелекту полягає в нереалістичних очікуваннях. Незважаючи на те, що машинне навчання — є перш за все інструментом, який дозволяє машинам вчитися на основі даних, публічний дискурс часто помилково приписує машинам здатність до людського мислення. Необхідно відрізнити міфи та наукову фантастику від емпіричної реальності [50].

Дискурс навколо штучного інтелекту часто зосереджується на його потенціалі покращувати різні аспекти людського життя. Він може чудово справляється з певними завданнями, перевершуючи людські здібності та, таким чином, покращуючи загальну якість життя. Штучний інтелект поступово змінює людський спосіб життя та має конкретні переваги [50].

Розглядається можливість майбутнього, в якому ШІ легко керуватиме завданнями, подібно до особистого секретаря. Наприклад, процес купівлі авіаквитків онлайн часто включає послідовність тривалих дій, включаючи вибір напрямків, розкладів і класів місць, а також порівняння цін на різних платформах. Незважаючи на те, що в поточних реалізаціях можуть бути недосконалості, постійне вдосконалення здібностей ШІ передвіщає майбутнє, у якому ШІ оптимізує завдання та усувуне пов’язані з цим труднощі для користувачів [50].

Інша перевага систем штучного інтелекту має значний вплив на подолання мовних бар’єрів, забезпечуючи трансформаційне рішення традиційних проблем вивчення мови, які характеризуються трудомісткими методами. У зв’язку з цим новаторські технології штучного інтелекту, такі як

Google Translate і пристрої голосової активації, такі як Google Home, стають інноваційними інструментами, які сприяють ефективній мовній комунікації. Інновації, керовані штучним інтелектом, проектують ландшафт, у якому досягнення в області штучного інтелекту сприяють досягненню можливостей мовного перекладу в реальному часі, які будуть легко інтегровані в компактні портативні пристрої. ШІ усуне часові обмеження, які супроводжують звичайні практики вивчення мови, що послужить початком зміни парадигми в оволодінні мовою. Перекладач Google використовує потужність алгоритмів машинного навчання, щоб забезпечити швидкий і точний переклад у різних лінгвістичних областях. Це полегшує вивчення мови та демократизує доступ до інформації, долаючи мовні бар'єри та сприяючи глобальному зв'язку. Пристрої з голосовою активацією, такі як Google Home, є прикладом трансформаційного впливу ШІ на мовну взаємодію. Завдяки обробці природної мови та машинному навчанню ці пристрої дозволяють людям взаємодіяти з технологіями бажаною мовою, долаючи мовні бар'єри та покращуючи взаємодію з користувачем. Цей прогрес не тільки спрощує спілкування, але й створює основу для більш інклюзивної та доступної цифрової взаємодії. Заглядаючи вперед, дискусії щодо мовних бар'єрів у сфері штучного інтелекту передбачають майбутнє, де мовний переклад у реальному часі буде бездоганно інтегрований у повсякденне життя. Концепція портативних пристроїв, оснащених розширеними можливостями штучного інтелекту, ілюструє перетин технологічних досягнень і дизайну, орієнтованого на користувача. Використання компактних пристроїв на базі штучного інтелекту для мовного перекладу в реальному часі окремими людьми обіцяє світ, де мова більше не буде перешкоджати ефективній комунікації, сприяючи більш взаємопов'язаній і гармонійній глобальній спільноті [50].

У сфері дослідження раку штучний інтелект пропонує багатообіцяючий, хоча й обмежений підхід. Здатність штучного інтелекту виявляти кореляції між генними мутаціями та впливом ліків від раку, особливо в областях, які

вислизають від спостереження людини, має незаперечний потенціал. Однак, переважаючі методи машинного навчання піддаються сумніву, коли справа доходить до виявлення причинно-наслідкових зв'язків, що призводить до сумнівів щодо їх медичного значення та впливу на покращення догляду за пацієнтами. Однак штучний інтелект демонструє практичну користь у конкретних медичних сценаріях, зокрема у візуальній діагностиці. У дерматології ШІ допомагає дерматопатологам досліджувати структурні зразки, розвиваючи столітній процес морфологічного аналізу. Історії успіху, зокрема партнерство між алгоритмами глибокого навчання та дерматологами у виявленні раку шкіри, підкреслюють досвід штучного інтелекту в розпізнаванні образів і його здатність досягати чудових результатів у конкретних ситуаціях. Хоча штучний інтелект забезпечує цінну діагностичну підтримку, життєво важливо розвіяти помилкове переконання, що він може повністю замінити медичних працівників. Роль штучного інтелекту полягає у створенні гіпотез, а не в заміні експертів. Цей процес має вирішальне значення для встановлення причинно-наслідкових зв'язків. Можливість інтерпретації рішень штучного інтелекту є все більшою сферою досліджень із різними застосуваннями. Розуміння стандартів для прийняття рішень ШІ має вирішальне значення для інформованої оцінки. Потенціал штучного інтелекту щодо розпізнавання образів сприяє прогресу в аналізі зображень, зокрема в радіології, дерматології та патології, хоча залишається проблемою гарантувати, що алгоритми встановлять розумні критерії для своїх завдань. Дослідники, такі як патологоанатоми, відзначили, що ШІ виявив раніше непомічені кореляції в медичних даних. Інтерпретоване машинне навчання дозволяє дослідникам визначати відповідні тканинні фактори, які корелюють з клінічними даними, надаючи інформацію про потенційні зв'язки. Проте проблеми залишаються, особливо в отриманні великої кількості навчальних даних для штучного інтелекту, які часто включають приватні дані про здоров'я, викликає етичні занепокоєння щодо власності та використання даних

приватними організаціями. У прагненні викорінити рак за допомогою штучного інтелекту доступність даних є значною перешкодою. Мета об'єднання даних про рак з усього світу заважає нерівномірному розподілу цієї інформації. Хоча інтеграція штучного інтелекту в дослідження раку є багатообіцяючою, надзвичайно важливо зосередитися на вирішенні етичних питань і забезпеченні прозорості в процесі прийняття рішень ШІ [59].

Інша сфера в якій системи ШІ можуть бути корисиними — це психотерапевтичне лікування, а саме інтеграція чат-ботів зі штучним інтелектом для оптимізації психотерапевтичних втручань і підвищення якості онлайн-психологічної підтримки. Інтеграція ШІ в терапевтичну практику, а саме, у консультуванні людей, які борються з розладами психічного здоров'я має значну перевагу [41].

Для відкриття нових матеріалів, дослідники все частіше використовують штучний інтелект для навігації у величезних базах даних і прогнозування характеристик можливих кристалічних структур. Розгалужена область хімії з її необмеженими комбінаціями атомів і численними матеріальними нейронними мережами створює труднощі у встановленні міцних і цінних зв'язків. Алгоритми ШІ представляють нову методологію, яка допомагає хімікам швидко розпізнавати перспективні сполуки для додаткових лабораторних експериментів. Хіміки Університету штату Флорида Кевін Раян, Джефф Ленгіле та Майкл Шатрук провели оцінку, у якій за допомогою нейронної мережі досліджували понад 30 000 встановлених кристалічних утворень неорганічних сполук. Алгоритм, якому не вистачало певної хімічної інформації, вивчив зв'язок елементів усередині кристалів виключно з геометричного розташування атомів. Навчаючи програму встановленим структурним типам, дослідники використовували штучний інтелект для оцінки ймовірності того, що окремі елементи займають визначені позиції сітки. Навчена програма штучного інтелекту виявила шаблони в даних, які вказують на те, що елементи з однієї групи в періодичній таблиці мають

схожість. Хоча явних знань про хімічні закони не було, створені правила для ШІ все ще дозволяли дослідникам оцінювати гіпотетичне існування випадково зібраних твердих тіл. Випробовуючи невідомі сполуки з експериментально підтвердженими кристалічними структурами, ШІ зміг визначити відомі структури серед першої десятки найімовірніших зв'язків, приблизно в 30% випадків. Незважаючи на те, що програма штучного інтелекту не гарантує повної точності прогнозування стабільних нових речовин, вона є важливим інструментом для створення списку численних можливих сполук. Ця методологія сумісна зі стандартними комп'ютерами та дозволяє швидко переглядати матеріали для дослідників, які шукають нові відкриття [22].

Важливо зауважити, що машини не роблять суджень так, як це роблять люди. Їхній процес прийняття рішень передбачає моделювання та вдосконалення рішень на основі заздалегідь запрограмованих цілей. Незважаючи на те, що вони пропонують перевагу швидкої обробки інформації, існують етичні проблеми при передачі обов'язків прийняття рішень комп'ютерам. Відповідальність за прийняття рішень має бути чітко визначена. Основний прогрес у штучному інтелекті був досягнутий завдяки штучним нейронним мережам із використанням статистичних інструментів. Однак розуміння процесів прийняття рішень у складних мережах створює проблеми через нездатність машин сформулювати обґрунтування своїх рішень. Враховуючи, що транспортні засоби, керовані програмним забезпеченням, спричиняють менше аварій, ніж ті, якими керує людина, чи виправдано віддавати пріоритет автономному водінню? Прикладом складності алгоритмічного прийняття рішень є автономне водіння. Хоча мета полягає в мінімізації шкоди від нещасних випадків, етичні обмеження обмежують алгоритмічні рішення та викликають питання про пріоритетність одного життя над іншим. Концепція «відокремлення людей» підкреслює етичні міркування в оптимізації комп'ютерних систем, зокрема в балансі індивідуального добробуту та суспільної вигоди. Які рішення мають бути делеговані

алгоритмам, а які потребують втручання людини? Алгоритмічні поради підходять для практичного вибору робочого місця, як-от вибір страхових полісів, тоді як екзистенційні рішення щодо кар'єри, стосунків і планування сім'ї вимагають людської проникливості. Незалежність у прийнятті таких рішень культивує більш глибокі стосунки та почуття особистої відповідальності за наслідки [38].

Вольфганг Губер, єпископ та голова Ради євангельської церкви в Німеччині, окреслює контури «етики цифровізації». Він наводить «принцип відповідальності», який філософ Ганс Йонас сформулював в однойменній книзі 1979 року: «Дій так, щоб наслідки твоїх дій були спроможними». сумісні з постійністю реального людського життя на землі». Губер хотів би, щоб ця сентенція застосовувалася не лише до поточної екологічної кризи, але й до індивідуальних і політичних дій у епоху комп'ютерної революції. Єпископ радить не приписувати комп'ютерам людські чи навіть надлюдські здібності, не дивлячись на те, що штучний інтелект неймовірно швидко вирішує поставлені завдання, він не може конкурувати з людськими здібностями, коли справа стосується саморозуміння та творчості. Він обговорює цифрову етику виключно з точки зору того, що щоденне використання складних машин означає для образу життя людини. Він відповідає на запитання як християнин: він відкидає утопії, які змагаються з Богом, обіцяючи людям безсмертя чи інші надлюдські здібності. Губер наголошує, що у розпал цифрової ери важливо зберігати ясність розуму, комп'ютери не замінять повністю людей, і не перетворять людей на богів [45].

Однак попри переваги використання штучного інтелекту, його помилки мають наслідки для реальних людей. Етичні наслідки абстрактної «чорної скриньки» вимагають ретельного вивчення. Аналіз відповідальності залучених сторін, прозорості та базових процесів обробки даних розкриває складні етичні основи, властиві системам ШІ. Бездумне використання даних, що ґрунтується на помилкових припущеннях, зроблених науковцями з

обробки даних, може призвести до дискримінаційних систем штучного інтелекту, зокрема в аналітиці правоохоронних органів, увічнюючи несправедливе упередження щодо певних етнічних меншин. Потенційна шкода, спричинена ненавмисним вибором поганих наборів даних для нейронних мереж, ілюструється гіпотетичними сценаріями, коли штучний інтелект, отримавши упереджену інформацію про певні групи, може зіткнутися з моральними труднощами. Вплив таких рішень, особливо в ситуаціях життя і смерті, вимагає етичних міркувань і навмисного відбору даних [50].

Вирішення етичних проблем штучного інтелекту, особливо в обробці природної мови, має вирішальне значення для забезпечення справедливості, прозорості та надійності. Старший науковий співробітник Microsoft Research у Нью-Йорку та ад'юнкт-професор Массачусетського університету Ханна Уоллах підкреслює суспільний вплив і відповідальність, пов'язану з обробкою природної мови. Основне етичне занепокоєння полягає в автоматичному відтворенні упереджень, які існують у навчальних даних, які використовуються для обробки керованої даними машинної мови. Уоллах цитує дослідження, яке демонструє, що мовні системи, навчені на таких джерелах, як газетні статті, мають тенденцію зміцнювати гендерні стереотипи. Наприклад, упередження, пов'язані між словами «медсестра» та «жінка», зберігаються в цих системах [60].

Прикладом автоматичного відтворення упереджень слугує програми Tay та Lensa. Корпорація Майкрософт у 2016 року представила прогруму Tay, на перший погляд невинний чат-бот, призначений для спілкування та вивчення взаємодії користувачів, особливо в контексті підліткової мови. Всупереч своїм доброзичливим намірам, Tay швидко продемонстрував дивовижну здатність вчитися та імпровізувати, використовуючи расистську, женоненависницьку та антифеміністичну мову під час політичних дискусій у соціальних мережах. Microsoft негайно видалив Tay з онлайн-платформ протягом дня після його

запуску [50]. Додаток Lensa для створення портретів демонструє сексистські тенденції. Концептуальна основа Lensa обертається навколо користувачів, які надсилають звичайні селфі, які потім піддаються алгоритмам штучного інтелекту, у результаті людина отримує художній портрет. Однак, зображення жінок ШІ характеризує сумнівними образами, переважно показуючи осіб, одягнених у облягаючий одяг, підкреслений широким оголенням грудей, і іноді представлений у стані повної оголеності. І навпаки, чоловіки в Lensa, як правило, зображуються в супергероїчних позах або прикрашені в костюмах з високим горлом. Важливо те, що такі програми постійно оновлюються новими зображеннями, отриманими з Інтернету, без явної згоди або відома зображених осіб і без будь-яких прав для тих, хто постраждав. Такі програми з використанням ШІ зміцнюють шкідливі стереотипи [56].

Ханна Уоллах наголошує на проблемі прихованих упереджень у системах машинного навчання, де минулі упередження ненавмисно зберігаються. Важливість аналізу помилок для розуміння реальних наслідків і те, що знання про загальну точність моделі є недостатнім. Також треба чітке визначення, як неточності впливають на різні демографічні показники, особливо розрізняючи модель, яка може похвалитися рівнем точності 95% для всіх демографічних показників, і ту, яка досягає рівня точності 100% для однієї групи, але значно поступається іншим. Уоллах усвідомлює складність подолання упереджень і наголошує на важливості аналізу помилок для людей, які стикаються з реальними наслідками. Хоча зусилля великих технологічних компаній щодо вирішення етичних проблем це позитивний крок, однак, складність кодування етичних алгоритмів залишається серйозною проблемою навіть для великих компаній, таких як Google. Етичні міркування в ШІ є складними і вимагають постійної уваги технологічної індустрії [60].

Сьогоднішньою етичною та технологічною проблемою є розповсюдження зображень, які називаються «діпфейк» (Deepfake). Складна взаємодія між програмами штучного інтелекту, які створюють реалістичні

зображення, та алгоритмами, призначеними для викриття таких підробок, становить складну ситуацію. Широке розповсюдження переконливих діпфейків про таких відомих діячів, як Папа Римський Франциск, Дональд Трамп, привернуло міжнародну увагу. Ці надзвичайно реалістичні зображення, створенні програмами штучного інтелекту, такими як «Dall-E», «Stable Diffusion» або «Midjourney», підкреслюють швидкий прогрес машинно створених зображень, які створюють дезінформацію. Розробка програмного забезпечення для створення діпфейків та алгоритмів його виявлення, створила конкурентне середовище [13].

Оновлення програмного забезпечення Midjourney у березні 2023 року, яке покращує якість діпфейків, демонструє безперервну еволюцію створення зображень, створених штучним інтелектом, і пов’язані проблеми з їх виявленням. Однак дослідники активно розробляють складні стратегії, які включають дослідження біомедичної інформації, як-от кровотік обличчя, або навчання програм штучного інтелекту розпізнавати ознаки діпфейків

Тим не менш, розкриття дипфейків все ще є великою проблемою. Будь-який успішний алгоритм, який використовується для виявлення підробок, може бути зловживаний тим самим ШІ, який він прагне виявити. У 2021 році відбувся значний прогрес у генеративному штучному інтелекті зі злиттям дифузійних і мовних моделей. Ця інноваційна техніка використовує короткий письмовий опис, щоб керувати підходом дифузії, забезпечуючи більший контроль над результатом, створеним ШІ. Такі компанії, як OpenAI, усвідомлюючи можливість неправомірного використання їхньої технології, інтегрували фільтри у своє програмне забезпечення, щоб запобігти створенню оманливого вмісту, особливо того, що стосується відомих осіб або непристойних матеріалів. Незважаючи на це, велика доступність генеративних алгоритмів з відкритим вихідним кодом викликає занепокоєння щодо нефільтрованого доступу до контенту, створеного ШІ. Удосконалення дифузійних моделей за останні роки вказує на їхню зростаючу ефективність у

створенні автентичних зображень. Оскільки час, необхідний для створення, зменшується, а результати стають переконливішими, важливість ефективних механізмів виявлення стає все більш актуальною [13].

Розпізнати діпфейки складно, але деякі ознаки можна помітити. Массачусетський технологічний інститут долучився до цього дискурсу, пропонуючи поради щодо розпізнавання зображень, створених штучним інтелектом, підкреслюючи важливість ретельного вивчення рис обличчя, тіней і елементів контексту. У міру розвитку механізмів виявлення, відрізнити реальні зображення від створених штучним інтелектом зображень залишається складною справою. Дослідження, в яких взяли участь понад 15 000 учасників, демонструють, що людське око, яке часто має більший досвід у виявленні тонких нюансів, ніж алгоритми, перевершує штучний інтелект у виявленні діпфейків. Meta запустила Deepfake Detection Challenge у 2019 році, надаючи конкурентоспроможну платформу для створення стійких алгоритмів виявлення. Незважаючи на докладені зусилля, програмне забезпечення досягло тільки 65 відсотків точності, проливаючи світло на складність проблеми виявлення. Отже, з розвитком цих технологій важливо створити ефективні методи виявлення діпфейків, щоб запобігти потенційним зловживанням і етичним проблемам [13].

В останні роки було докладено значних зусиль, спрямованих на розробку алгоритмів штучного інтелекту для точного розрізнення автентичного вмісту від маніпуляційного. Згорткові нейронні мережі, спеціалізовані алгоритми, натхненні зоровою корою ссавців, використовуються для навчання програм штучного інтелекту розрізняти реальні зображення від створених ШІ. Однак виникає серйозна проблема, оскільки ці алгоритми штучного інтелекту функціонують як чорні ящики, пропонуючи мало розуміння того, як вони визначають автентичність, навіть якщо вони ефективні. У результаті дослідники вивчають альтернативні підходи для автоматичного визначення дипфейків, зокрема у відео.

Розробка нових методологій може призвести до незрозумілих результатів у існуючих алгоритмах. Програма «FakeCatcher» була представлена в 2020 році дослідниками Університету штату Нью-Йорк, використовуючи фотоплетизмографію, метод, який аналізує зміни кольору шкіри для визначення кровотоку. Програма досягла багатообіцяючих результатів у відрізненні справжніх відео від створених штучним інтелектом. Так само програма PhaseForensics, створена дослідниками Каліфорнійського університету в Санта-Барбарі, використовує машинне навчання та біомедичні маркери з акцентом на рухах губ, щоб визначити, чи є вміст природним чи створеним штучним інтелектом [13].

Симбіотичний зв'язок між алгоритмами виявлення та штучним інтелектом, що створює зображення, створює парадоксальну динаміку, незважаючи на прогрес. Програми для створення зображень покращують свою реалістичність шляхом повторного налаштування параметрів на основі зворотного зв'язку програмного забезпечення виявлення. Однак значна різниця в ресурсах, доступних для розробки програм для створення синтетичних зображень, і для програм виявлення викликає занепокоєння.
У відповідь на цей виклик дослідники пропонують змінити парадигму, виступаючи за те, щоб програми ШІ маркували свої продукти унікальним водяним знаком. Цей підхід водяних знаків призначає числові значення пікселям, створюючи тонкий візерунок, який відрізняє створений штучним інтелектом вміст від автентичного. Хоча цей водяний знак можна видалити, цей процес є стримуючим фактором, вимагаючи від користувачів змінювати основний вихідний код [13].

Альтернативним методом є використання криптографічних підписів, які призначають унікальний підпис справжнім записам, роблячи їх захищеними від фальсифікації або підробки. Ця техніка забезпечує сильну гарантію автентичності файлу, оскільки його зміна призведе до миттєвої недійсності криптографічного підпису. Впровадження водяних знаків, спеціально

розроблених для зображень, створених штучним інтелектом, разом із використанням криптографічних сертифікатів для автентичних записів представляє багатогранний підхід до значного зменшення поширення діпфейків. Використання криптографічних підписів діє як захисний засіб від зловмисних змін, зміцнюючи цілісність цифрового вмісту. Цей двокомпонентний підхід покращує процес автентифікації та покращує захист від шахрайського контенту, створеного ШІ. Ці вдосконалені методи використовуються для підтримки надійності та точності цифрових записів, одночасно пом'якшуючи потенційну шкоду від оманливих маніпуляцій за допомогою технологій штучного інтелекту [13].

Швидкий прогрес машинного навчання підштовхнув рух за тимчасову зупинку навчання потужних моделей ШІ. Відомі підписанти відкритого листа, такі як Джошуа Бенгіо та Ілон Маск, пропагують обережний підхід, наголошуючи на важливості оцінки потенційних переваг і керованих ризиків. Такі ініціативи, як обов'язкове маркування та криптографічні сертифікати, були запропоновані як початкові кроки до законодавчого регулювання та для навігації в етичному ландшафті зображень, створених штучним інтелектом [13].

Інша проблема у цифрову епоху — це зменшення робочих місць через уявлення про те, що технології зроблять багато професій застарілими. Професор Стефан А. Янсен, голова центру благодійності та громадянського суспільства, кидає виклик цим тривогам, висвітлюючи історичні закономірності технологічної еволюції та появу нових робочих можливостей. Підкреслюючи динамічний характер технологічного прогресу, Янсен пояснює, що історично кожен технологічний розвиток замінював відповідні професії. Він скептично ставиться до досліджень, які передбачають втрату робочих місць, обґрунтовуючи це через методологічні недоліки у цих дослідженнях. Тому що вплив цифровізації на робочі місця є багатогранним і соціально складним [19]. Хоча дослідження Deloitte та Оксфордського

університету 2014 року стверджує, що 35% робочих місць у Великобританії знаходяться під загрозою. Індустрія інформаційних технологій вартістю 150 мільярдів доларів, в якій працюють понад чотири мільйони людей, зіткнулася зі зміною парадигми з появою штучного інтелекту та автоматизації як потенційних факторів зайнятості. Однак занепокоєння є великим: прогнози свідчать про втрату до 69% робочих місць в Індії та 77% у Китаї через інтеграцію ШІ. Очікувані наслідки інтеграції ШІ, подібні до історичних технологій загального призначення, таких як паровий двигун, електрика, двигун внутрішнього згоряння або мікропроцесор, припускають заміну певних завдань, які виконує людина. Прискорення технологічного прогресу викликає занепокоєння щодо витіснення робочих місць, що вимагає вжиття проактивних заходів. Замість того, щоб чинити опір технологічному прогресу через регулювання чи оподаткування, більш конструктивний підхід передбачає покращення освіти, сприяння універсальності та створення мережі соціального захисту для підтримки людей у адаптації до мінливих робочих ландшафтів [50].

Професор Янсен критикує німецьку систему освіти, заявляючи, що її зосередженість на традиційних професіях робить її погано підготовленою до викликів, пов’язаних із цифровізацією. Він підтримує освітні підходи, які розвивають креативність, критичне мислення та навички вирішення проблем, підкреслюючи важливість підготовки людей до нестандартних процедур і сценаріїв прийняття рішень, які неможливо автоматизувати. Важливость адаптації освіти до мінливих вимог ринку праці та відхід віл традиційних моделей освіти. Звертаючись до когнітивного аспекту взаємодії людини та машини, Янсен наголошує на необхідності досягнення суспільством когнітивної переваги над машинами. Він також передбачає новий вид вищої освіти, який зосереджується на етичних міркуваннях, нормативних аспектах і соціально-технологічному розширенні свідомості, позиціонуючи цікавість як рушійну силу в цю швидко розвиваючу цифрову еру [19].

Оскільки інтелект — це цілий набір рішень для незалежних проблем, немає причин боятися раптової появи надлюдської машини, яка мислить, хоча завжди краще звертати увагу на потенційний ризик. Звичайно, кожна з багатьох технологій, яка з'являються для вирішення людських проблем, ймовірно, сама по собі є потужною — і, отже, потенційно небезпечною у її неправильному використанні, як і більшість технологій. Таким чином, повинні бути вжиті відповідні заходи безпеки у по'єднанні з етичними принципами. Крім того, існує потреба в постійному моніторингу, можливо, незалежною багатонаціональною рганізацією [15].

Наразі Німеччина працює над створенням штучного інтелекту, під назвою Open GPT-X, яка має включати європейські цінності і захист даних. «Важливо, щоб ми в Європі розробляли таку технологію, яка базується на наших цінностях і потребах і відповідає вимогам захисту даних», — говорить Делара Бурхардт німецька політична дічка, член Соціал-демократичної партії Німеччини, депутат Європейського парламенту. «Ми втратимо свій суверенітет, якщо передамо всі наші дані, особливо внутрішні дані компанії, американській компанії» [14].

Генеральна конференція Організації Об'єднаних Націй з питань освіти, науки і культури (ЮНЕСКО), на своїй 41-й сесії у 2021 році представила рекомендації щодо етики штучного інтелекту. Важливою частиною рекомендацій є дотримання конкретних цінностей і принципів протягом усього життєвого циклу системи штучного інтелекту і його повинні підтримувати всі залучені організації. Просування цих цінностей можна досягти шляхом перегляду чинних законів, нормативних актів і керівних принципів ведення бізнесу, щоб вони відповідали міжнародному праву, включаючи Статут Організації Об'єднаних Націй, зобов'язання держав-членів щодо прав людини та погоджені на міжнародному рівні [51].

Ключові принципи та цінності ЮНЕСКО щодо етики штучного інтелекту:

1. Непорушна та притаманна кожній людині гідність є наріжним каменем універсальної системи прав людини та основних свобод. Отже, основний імператив полягає в безперервній повазі, захисті та заохоченні людської гідності та прав, як це встановлено міжнародним правом, включаючи сферу міжнародного права прав людини, протягом усього життєвого циклу систем ШІ. Людська гідність, що охоплює внутрішню та однакову цінність кожної людини, незалежно від різноманітних якостей, вимагає визнання без шкоди для раси, кольору шкіри, походження, статі, віку, мови, релігії, політичних поглядів, національного походження, етнічного походження, соціального походження, економічного становища або соціальні умови народження, інвалідність та інші відповідні підстави;
2. Процвітання навколишнього середовища та екосистем необхідно визнавати, захищати та розвивати протягом усього життєвого циклу систем ШІ. Крім того, навколишнє середовище та екосистема є фундаментальними передумовами для того, щоб людство та інші живі істоти могли пожинати переваги прогресу ШІ. Усі учасники життєвого циклу систем штучного інтелекту зобов’язані дотримуватися відповідного міжнародного права та національного законодавства, а також встановлених стандартів і практик, таких як запобіжні заходи, призначені для захисту та відновлення навколишнього середовища та екосистем, а також сприяння сталому розвитку. Для них вкрай необхідно зменшити вплив систем ШІ на навколишнє середовище, включаючи, але не обмежуючись, їхній вуглецевий слід;
3. Забезпечення поваги, захисту та заохочення різноманітності та інклюзивності є обов’язковими протягом усього життєвого циклу систем штучного інтелекту відповідно до міжнародного права, включаючи право людини. Цього можна досягти шляхом активного заохочення участі всіх осіб або груп, незалежно від раси, кольору шкіри, походження, статі, віку, мови, релігії, політичних поглядів, національного походження, етнічного

походження, соціального походження, економічного чи соціального стану народження, інвалідність та інші підстави;

4. Учасники штучного інтелекту повинні активно брати участь у сприянні розвитку мирних і справедливих суспільств, заснованих на взаємопов'язаному майбутньому, яке приносить користь усім і відповідає правам людини та основним свободам. Внутрішня цінність проживання в суспільствах, що характеризуються миром і справедливістю, підкреслює потенціал систем штучного інтелекту протягом усього життєвого циклу робити всебічний внесок у взаємозалежність усіх живих істот один з одним і з природним середовищем;
5. Необхідно визнати, що технології ШІ самі по собі не гарантують процвітання людей, навколишнього середовища та екосистем. Крім того, усі процеси, пов'язані з життєвим циклом систем штучного інтелекту, не повинні перевищувати те, що є важливим для досягнення законних цілей або завдань, і вони повинні відповідати контексту. У разі потенційної шкоди людям, правам людини та основним свободам, громадам, суспільству в цілому або навколишньому середовищу та екосистемам надзвичайно важливо запровадити процедури оцінки ризику та вжити заходів для запобігання такій шкоді;
6. Учасники штучного інтелекту зобов'язані відстоювати соціальну справедливість, забезпечуючи справедливість і відсутність дискримінації відповідно до міжнародного права. Це вимагає інклюзивної стратегії, щоб зробити переваги технологій штучного інтелекту загальнодоступними, враховуючи різноманітні потреби різних вікових груп, культурних систем, мовних спільнот, людей з обмеженими можливостями, дівчат і жінок, а також тих, хто є знедоленими, маргіналізованими чи вразливими;
7. Необхідно активно уникати небажаної шкоди, такої як ризики безпеці та вразливості до атак, і боротися з ними протягом усього життєвого циклу систем ШІ, щоб гарантувати безпеку для людей, навколишнього

середовища та екосистем. Досягнення безпечного та захищеного штучного інтелекту залежить від створення стійких структур доступу до даних із захистом конфіденційності, які сприяють вдосконаленому навчанню та перевірці моделей штучного інтелекту шляхом використання високоякісних даних;

8. Конфіденційність, яка розглядається як невід'ємне право для захисту людської гідності, незалежності та свободи волі, має підтримуватися nf захищатися протягом усього життєвого циклу систем ШІ. Збір, використання, обмін, архівування та видалення даних для систем штучного інтелекту мають відповідати міжнародному праву та дотримуватися цінностей і принципів, викладених у цій Рекомендації ЮНЕСКО, а також поважати відповідні національні, регіональні та міжнародні правові рамки;
9. Прозорість і зрозумілість систем штучного інтелекту часто є фундаментальними передумовами для забезпечення поваги, захисту та заохочення прав людини, основних свобод і етичних принципів. Прозорість є необхідною для ефективної роботи відповідних національних і міжнародних режимів відповідальності. Відсутність прозорості може підірвати здатність оскаржувати рішення, засновані на результатах, створених системами штучного інтелекту, потенційно порушуючи право на справедливий суд, і обмежуючи правові сфери, в яких ці системи можуть використовуватися [51].

Рекомендація ЮНЕСКО щодо етики штучного інтелекту є значним внеском у глобальну структуру цифрових прав людини. Історично склалося так, що захист прав людини в контексті цифровізації багато держав переважно здійснювався в рамках національних нормативних рамок. Визнаючи нагальну потребу посилити захист прав людини в умовах цифрової трансформації, Федеральний уряд Німеччини визначив це питання пріоритетним у своєму Плані дій з прав людини на 2021–2022 роки. Рекомендація ЮНЕСКО відповідає різноманітному впливу ШІ на економічні, соціальні, культурні,

громадянські та політичні права, що прагне встановити принципи та правила, які використовують потенціал штучного інтелекту, одночасно захищаючи індивідуальні свободи та забезпечуючи соціальну згуртованість. Незважаючи на відсутність прямої згадки про права людини в контексті ШІ в Коаліційній угоді уряду Німеччини на 2021–2025 роки, визнається важливість ШІ як «сфери майбутнього». Законодавча пропозиція Європейської комісії щодо регулювання штучного інтелекту характеризується амбітним характером і має бути обов'язковою для держав-членів ЄС після прийняття. Проект Регламенту, представлений 21 квітня 2021 року, відображає важливий початковий крок у потенційно складному законодавчому процесі [27].

Висновки до розділу 3

Дослідження взаємозв'язку між штучним інтелектом і людським інтелектом та свідомістю вказують на історичну прогресію, позначену періодами очікування, що чергуються з невдачами або «зимами ШІ». Складність визначення інтелекту очевидна, і хоча Франсуа Шолле запропонував визначення, засноване на здатності отримувати навички, відсутність широко визнаного визначення підкреслює складність цього поняття. Незважаючи на значний прогрес у сфері ШІ, відтворення людського інтелекту за допомогою цих технології залишається недосяжним. Поточні дослідження нейробіологів надалі спрямовані на виявлення нейронного кореляту свідомості, необхідного для розвитку майбутнього ШІ.

Впровадження штучного інтелекту в різні сфери людської діяльності пов'язане як з перевагами, так і з етичними недоліками. Нереалістичні очікування, потенційні упередження та складність усунення неявних упереджень у системах машинного навчання підкреслюють важливість підзвітної розробки ШІ. Можливі переваги покращення мовного перекладу разом із медичними та науковими дослідженнями та психотерапевтичними

втручаннями необхідно збалансувати з потенційними занепокоєннями, такими як етичні наслідки, зокрема безробіття та поширення маніпуляційного контенту.

Щоб забезпечити об’єктивність, прозорість і надійність, вирішення етичних проблем, пов’язаних зі штучним інтелектом, особливо в області обробки природної мови, має першочергове значення. Встановлення етичних принципів, регулярний моніторинг і створення неупереджених систем штучного інтелекту мають вирішальне значення для життя людей. Рекомендації ЮНЕСКО з етики штучного інтелекту роблять цінний внесок у глобальну цифрову структуру прав людини, наголошуючи на важливості дотримання певних цінностей протягом усього життєвого циклу систем штучного інтелекту. Регламент Європейської комісії щодо штучного інтелекту знаменує важливу віху в законодавчому процесі, оскільки він встановлює стандарти для держав-членів ЄС.

Баланс між потенційними перевагами та етичними проблемами, такими як захист прав людини та сприяння відповідальним інноваціям, має важливе значення для навігації у сфері штучного інтелекту, що розвивається. Досягнення повного потенціалу штучного інтелекту для суспільної користі вимагає постійної міждисциплінарної співпраці, дотримання етичних принципів і міжнародного співробітництва.

ВИСНОВКИ

Траєкторія розвитку штучного інтелекту розгортається як стійкий і трансформуючий наратив, який перемежовується моментами очікування, невдач і відродження. Його еволюційний шлях, який бере свій концептуальний початок із стародавніх міфів і ґрунтується на піонерському внеску таких діячів, як Беббідж, Цузе і Тьюринг, відображає синтез технологічного прогресу, етичних дилем та еволюції суспільних перспектив. Взаємодія цих факторів суттєво впливає на траєкторію розвитку ШІ.

Поява глибокого навчання та нейронних мереж, натхненних складними структурами людського мозку, означає ключову зміну парадигми в еволюції ШІ. Хоча в алгоритмах машинного навчання було досягнуто значних успіхів, неймовірна складність людського мозку залишається серйозною проблемою для створення цифрових аналогів. У міру розвитку суспільства стає критично важливим знайти баланс між потенційними перевагами та ризиками штучного інтелекту. Етичні міркування, підкреслені прозорістю, підзвітністю та розумінням впливу ШІ на суспільство, повинні керувати траєкторією досліджень та розробки систем ШІ.

Поточні дослідження складності біологічних нейронів відкривають чудові можливості для створення альтернативних алгоритмів і покращення майбутнього розвитку технології ШІ. Також, заслуговує на увагу спостереження про те, що здатність до навчання п’ятирічних дітей перевершує можливості найпотужніших комп’ютерів, підкреслює постійну потребу вдосконалення та досліджень у цій галузі, що швидко розвивається.

Комплексний аналіз сучасних моделей штучного інтелекту, включаючи ChatGPT, DALL-E, Minerva, Lamda, Google Bard і AzureAI, дає розуміння поточного ландшафту цієї галузі. Незважаючи на надзвичайний прогрес моделей, такі проблеми, як помилки, непрозорість, упередженість і дебати щодо оптимального розміру моделі залишаються.

Пріоритет етики є життєво важливим серед стрімкого технологічного прогресу. Мінімізація ризиків і забезпечення етичного використання вимагає відповідальних практик розробки, прозорого спілкування та превентивних заходів для пом'якшення упереджень системи ШІ. Щоб повністю розкрити потенціал штучного інтелекту, міждисциплінарна співпраця, непохитне дотримання етичних принципів і прагнення до мінімізації впливу на навколишнє середовище мають вирішальне значення.

Складний зв'язок між штучним інтелектом, людським інтелектом і свідомістю проявляється як багатогранна подорож із нюансами, що перемежовується фазами очікування та невизначеності. Завдання визначення інтелекту виявилося надзвичайно складним викликом, і відтворення людського інтелекту за допомогою технологій штучного інтелекту залишається неможливою через фундаментальну відмінність в архітектурі біологічних нейронів і їхніх цифрових аналогів, а також загадковій природі людської свідомості.

Впровадження штучного інтелекту в різні аспекти людського існування несе як переваги, так і проблеми. Відповідальна розробка та впровадження нормативних актів має важливе значення для пом'якшення проблем із потенційними наслідками для окремих осіб та суспільства в цілому. Вирішення етичних дилем, особливо в контексті обробки природної мови, є вирішальним кроком у гарантуванні справедливості, прозорості та надійності систем ШІ.

У нинішній еволюції галузі, вкрай важливо знайти баланс між потенційними вигодами та етичними дилемами. Щоб використовувати трансформаційний потенціал штучного інтелекту на користь людства та вирішення етичних викликів, необхідні стійке міждисциплінарне співробітництво, дотримання етичних норм і міжнародна співпраця. При впровадженні етичних принципів у розробку штучного інтелекту етичні міркування є надзвичайно важливими. Забезпечення відповідності суспільним

та індивідуальним цінностям має вирішальне значення з розвитком технологій. Надійна інтеграція штучного інтелекту в людське буття залежить від встановлення пріоритетів етичних стандартів.

Список використаних джерел:

1. Філософський енциклопедичний словник / ред. В. Шинкарук. Київ : Абрис, 2002. 742 с.
2. Alan Turing. *Wikipedia – Die freie Enzyklopädie*. URL: https://de.wikipedia.org/wiki/Alan_Turing (дата звернення: 10.10.2023).
3. Al-Khalili J. What's next? Even scientists can't predict the future or can they?. London : CPI Group (UK) Ltd, 2017. 236 c.
4. Altman S. ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. it's a mistake. *Twitter*. URL: https://twitter.com/sama/status/1601731295792414720?s=20 (дата звернення: 13.11.2023).
5. Analytical engine. *Wikipedia – Die freie Enzyklopädie*. URL: https://de.wikipedia.org/wiki/Analytical_Engine (дата звернення: 20.10.2023).
6. Ananthaswamy A. Informatik - mit KI das menschliche Gehirn verstehen. *Spektrum der Wissenschaft Kompakt - Mensch & Maschine*. 2022. № 21. С. 4–15.
7. Ananthaswamy A. Informatik - Programm mit Köpfchen. Spektrum der Wissenschaft Kompakt - Mensch und Maschine. 2022. № 21. С. 17–25.
8. Ananthaswamy A. Sprachmodelle - Ist bei einer KI größer immer besser?. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 37–48.
9. Antiquity - A history of artificial intelligence. *A History of Artificial Intelligence*. URL: https://ahistoryofai.com/antiquity/ (дата звернення: 20.09.2023).
10. ARC abstraction & reasoning corpus. *Lab42*. URL: https://lab42.global/arc/ (дата звернення: 17.11.2023).

11. Artificial intelligence. *Cambridge Dictionary*. URL: https://dictionary.cambridge.org/dictionary/english/artificial-intelligence (дата звернення: 06.12.2023).
12. Azure KI. *Microsoft*. URL: https://azure.microsoft.com/de-de/solutions/ai (дата звернення: 21.11.2023).
13. Bischoff M. Deepfakes - Wie lassen sich KI-generierte Bilder enttarnen?. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 58–69.
14. Bischoff M. Künstliche Intelligenz - Was steckt hinter ChatGPT & Co?. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 4–23.
15. Brockman J. What to think about machines that think: today's leading thinkers on the age of machine intelligence. Harper Perennial, 2015. 576 с.
16. Caliskan A., Bryson J. J., Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. 2017. С. 1657. URL: https://doi.org/10.1126/science.aal4230 (дата звернення: 13.10.2023).
17. Chollet F. On the measure of intelligence. Google, Inc, 2019. 64 с.
18. Dartmouth conference. *Wikipedia – Die freie Enzyklopädie*. URL: https://de.wikipedia.org/wiki/Dartmouth_Conference (дата звернення: 24.10.2023).
19. Digitalisierung - Was bleibt für den Menschen noch zu tun?. Spektrum der Wissenschaft Kompakt - Künstliche Intelligenz - Der Weg in die Anwendung. 2019. № 40. С. 41–45.
20. Electroencephalograph. *Cambridge Dictionary*. URL: https://dictionary.cambridge.org/dictionary/english/artificial-intelligence (дата звернення: 06.12.2023).
21. General problem solver. *Wikipedia – Die freie Enzyklopädie*. URL: https://de.wikipedia.org/wiki/General_Problem_Solver (дата звернення: 23.10.2023).

22. Groß M. Chemie - Künstliche Intelligenz entdeckt neue Stoffe. Spektrum der Wissenschaft Kompakt - Künstliche Intelligenz - Der Weg in die Anwendung. 2019. № 40. С. 17–21.
23. Haseeb Hassan M. Google AI – Minerva for quantitative reasoning problems. *Medium*. URL: https://medium.com/@TheHaseebHassan/google-ai-minerva-for-quantitative-reasoning-problems-4f81fa5a4b77 (дата звернення: 18.11.2023).
24. Hopffgarten A. Lernwelten - Maschinen das Träumen lehren. Spektrum der Wissenschaft - Künstliche Intelligenz - Wie Maschinen lernen lernen. 2018. № 39. С. 21–27.
25. How ChatGPT and our language models are developed. *OpenAI Help Center*. URL: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed#h_61e36f9199 (дата звернення: 01.11.2023).
26. Kerner S. M. What is Dall-E (Dall-E 2) and how does it work?. *Enterprise AI*. URL: https://www.techtarget.com/searchenterpriseai/definition/Dall-E (дата звернення: 17.11.2023).
27. Kettemann M. C. Recommendation on the ethics of artificial intelligence conditions for the implementation in Germany / ред. Deutsche UNESCO-Kommission. 48 с.
28. Kognitionswissenschaft - Was KI über unsere Intelligenz lehrt / A. Newen та ін. *Spektrum der Wissenschaft Kompakt - Mensch & Maschine*. 2022. № 21. С. 32–38.
29. Konrad Zuse. *Wikipedia – Die freie Enzyklopädie*. URL: https://de.wikipedia.org/wiki/Konrad_Zuse (дата звернення: 19.10.2023).
30. Krauß P., Maier A. Bewusstsein - Der Geist in der Maschine. *Spektrum der Wissenschaft Kompakt - Mensch und Maschine*. 2022. № 21. С. 39–50.
31. LaMDA: language models for dialog applications. *Google Research*. URL: https://research.google/pubs/pub51115/#:~:text=LaMDA%20is%20a%20f

amily%20of,dialog%20data%20and%20web%20text (дата звернення: 12.11.2023).

32. Logic theorist explained - everything you need to know. *History-Computer*. URL: https://history-computer.com/logic-theorist/ (дата звернення: 24.10.2023).

33. Lutkevich B. What is AI winter? Definition, history and timeline. *Enterprise AI*. URL: https://www.techtarget.com/searchenterpriseai/definition/AI-winter (дата звернення: 27.10.2023).

34. Newell A., Shaw J. C., Simon H. A. Report on a general problem-solving program. The rand corporation - Carnegie institute of technology, 1958. 29 с.

35. Nida-Rümelin J. Was ist gerecht?. Spektrum der Wissenschaft Highlights - Die zwölf größten Rätsel der Philosophie. 2022. № 2. С. 50–57.

36. Piantadosi S. *Twitter*. URL: https://twitter.com/spiantado/status/1599462405225881600 (дата звернення: 09.11.2023).

37. Plekat S.-M. Wieso diskriminieren künstliche Intelligenzen?. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 57.

38. Ramge T. Algorithmen - »Entscheiden kann nur der Mensch«. Spektrum der Wissenschaft Kompakt - Künstliche Intelligenz - Der Weg in die Anwendung. 2019. № 40. С. 52–57.

39. Ramge T. Entscheidungen - Management by Null und Eins. Spektrum der Wissenschaft Kompakt - Künstliche Intelligenz - Der Weg in die Anwendung. 2019. № 40. С. 46–51.

40. Ramge T. Intelligenz - Wie schlau ist künstliche Intelligenz. Spektrum der Wissenschaft Kompakt - Künstliche Intelligenz - Der Weg in die Anwendung. 2019. № 40. С. 30–33.

41. Retzbach J. Psychotherapie - Algorithmen als Hilfstherapeuten. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 55–56.
42. Roberts M. Google Bard statistics & facts. *MLYearning*. URL: https://www.mlyearning.org/google-bard-statistics-facts/ (дата звернення: 10.11.2023).
43. Schlicht T. Philosophie des Geistes - Dem Bewusstsein auf der Spur. Spektrum der Wissenschaft Highlights - Die zwölf größten Rätsel der Philosophie. 2022. № 2. С. 16–23.
44. Smith G. What does “fairness” mean for machine learning systems?. *Berkeley Haas*. URL: https://haas.berkeley.edu/equity/ (дата звернення: 29.11.2023).
45. Springer M. »Menschen, Götter und Maschinen« - Einen klaren Kopf behalten. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 71–72.
46. Sterzer P. Die Illusion der Vernunft. Berlin : Ullstein, 2022. 320 с.
47. Synaptic weight. *Wikipedia, the free encyclopedia*. URL: https://en.wikipedia.org/wiki/Synaptic_weight (дата звернення: 27.10.2023).
48. The internet’s new favorite AI proposes torturing iranians and surveilling mosques. *The Intercept*. URL: https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/ (дата звернення: 15.11.2023).
49. Thinking. *Cambridge Dictionary*. URL: https://dictionary.cambridge.org/dictionary/english/artificial-intelligence (дата звернення: 06.12.2023).
50. UNESCO. Human decisions thoughts on AI / ред.: B. Lasry, H. Kobayashi. Paris : Imprimerie Frag Paris, 2018. 137 с.
51. UNESCO. Recommendation on the ethics of artificial intelligence. France, 2021. 43 с.

52. van Berkel N., Sarsenbayeva Z., Goncalves J. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FACCT. *International journal of human-computer studies*. 2022. С. 102954. URL: https://doi.org/10.1016/j.ijhcs.2022.102954 (дата звернення: 30.11.2023).
53. Vandeput N. A brief history of neural networks. *Medium*. URL: https://medium.com/analytics-vidhya/a-brief-history-of-neural-networks-c234639a43f1 (дата звернення: 31.10.2023).
54. Vosgerau G. Kognition - Sprache und Denken. Spektrum der Wissenschaft Highlights - Die zwölf größten Rätsel der Philosophie. 2022. № 2. С. 63–67.
55. Whitten A. Tiefe Netzwerke - Wie komplex sind Neurone wirklich?. *Spektrum der Wissenschaft Kompakt - Mensch und Maschine*. 2022. № 21. С. 26–31.
56. Wie sexistisch ist die KI-Porträt-App Lensa?, 2023. *Made for minds*. URL: https://www.dw.com/de/wie-sexistisch-ist-die-ki-porträt-app-lensa/video-64882654 (дата звернення: 12.11.2023).
57. Winter D. Intentionalität - Warum KI nichts wollen kann. *Spektrum der Wissenschaft Kompakt - Mensch & Maschine*. 2022. № 21. С. 51–56.
58. Wolfangel E. ChatGPT - Das sprachgewaltige Plappermaul. Spektrum der Wissenschaft Kompakt - Künstliche Gespräche Kommunikation mit KI-Chatbots. 2023. № 19. С. 24–29.
59. Wolfangel E. Krebs und big Data - Lange gesund leben – dank KI. Spektrum der Wissenschaft Kompakt - Künstliche Intelligenz - Der Weg in die Anwendung. 2019. № 40. С. 66–70.
60. Wolfangel E. Vorteile und Klischees - Wo hat sie das nur gelernt?. Spektrum der Wissenschaft - Künstliche Intelligenz - Wie Maschinen lernen lernen. 2018. № 39. С. 62–65.