МІНІМАЛЬНИЙ ТЕСТ ТЮРІНГА: ПСИХОЛОГІЯ РОЗУМІННЯ ВІДМІННОСТЕЙ ЛЮДИНИ ТА МАШИНИ

MINIMAL TURING TEST: PSYCHOLOGY OF UNDERSTANDING DIFFERENCES BETWEEN HUMANS AND MACHINES

УДК 159.9
DOI https://doi.org/10.32782/2663-5208.2024.57.51

Кириченко В.В.
д.психол.н., доцент,
професор кафедри соціальної та практичної психології
Житомирський державний університет імені Івана Франка

У статті розкриваються соціально-психологічні особливості взаємодії людини із штучним інтелектом. Опираючись на оригінальні дослідження Алана Тюрінга, ми дійшли висновку, що під час взаємодії зі штучним інтелектом відбувається екстраполяція рис особистості щодо технологій штучного інтелекту. Це відбувається з метою пошуку відмінностей людини та машини, протиставлення їх властивостей, але на основі пошуку властивостей, які є суто людськими. У дослідженні Алана Тюрінга людина диференціює людину та машину на основі їх інтелектуальних дій, проте у XXI столітті цей тест було пройдено, що свідчить про значний розвиток технологій штучного інтелекту та їх широку інтеграцію у всі сфери діяльності особистості. Орієнтуючись на зразки популярної культури, ми дійшли висновку, що люди протиставляють себе штучному інтелекту, вбачають у ньому небезпеку для себе, а його дії оцінюють як загрозу. Метою нашого дослідження було встановити, чи прослідковується вплив антропоморфізму на атрибуції дій і результатів діяльності штучного інтелекту та якої модальності властивості приписують досліджувані штучному інтелекту. Ми запропонували досліджуваним оцінити, наскільки запропоновані риси можуть бути властивими для людини та штучного інтелекту: «доброта», «співпереживання», «чуйність», «працелюбність», «відповідальність», «підступність», «жорстокість», «брехливість», «безвідповідальність», «інфантильність». За результатами дослідження було встановлено, що досліджувані схильні до атрибуції рис особистості неживим об'єктам, що є свідченням впливу антропоморфізму на атрибуцію об'єктивних властивостей штучного інтелекту. Досліджувані приписують технологіям штучного інтелекту особистісні риси, які позначають їх технологічні властивості. Досліджувані приписують технологіям штучного інтелекту риси різної модальності: як позитивні, так і негативні, тому не підтвердилася гіпотеза про диференціацію властивостей людини та машини на основі атрибуції рис особистості різної модальності.

Ключові слова: *штучний інтелект, риси особистості, антропоморфізми, інформаційне суспільство.*

The article explores the socio-psychological aspects of human interaction with artificial intelligence. Drawing on the original research of Alan Turing, we conclude that during interaction with artificial intelligence, there is an extrapolation of personality traits regarding artificial intelligence technologies. This occurs to seek differences between humans and machines, contrasting their properties, but based on the search for properties that are purely human. In Turing's research, humans differentiated between humans and machines based on their intellectual actions, but in the 21st century, this test has been passed, indicating significant development in artificial intelligence technologies and their wide integration into all aspects of personal activity. By relying on examples from popular culture, we have concluded that people perceive themselves in opposition to artificial intelligence, seeing it as a threat and evaluating its actions as dangerous. The goal of our research was to determine whether anthropomorphism influences the attributions of actions and results of artificial intelligence activity, and which modality properties are attributed to the studied artificial intelligence. We asked the participants to evaluate how the proposed traits could be characteristic of both humans and artificial intelligence: "kindness", "empathy", "sensitivity", "industriousness", "responsibility", "deviousness", "cruelty", "dishonesty", "irresponsibility" and "infantilism". The research results revealed that the participants tend to attribute personality traits to inanimate objects, indicating the influence of anthropomorphism on attributing objective properties to artificial intelligence. The participants ascribe personality traits to artificial intelligence technologies, marking their technological properties. The participants attribute various modalities of traits to artificial intelligence technologies, both positive and negative, thus not confirming the hypothesis of differentiating human and machine properties based on the attribution of personality traits of different modalities.

Key words: *artificial intelligence, personality traits, anthropomorphisms, information society.*

Постановка проблеми. З моменту проведення експерименту Алана Тюрінга, відомого широкому загалу як «гра в імітацію», минуло понад 70 років. За цей час цифрові технології пройшли тривалий і складний шлях еволюції: від здатності імітувати людське мислення, емоції до імітації творчості. Цей шлях був не однорідним і супроводжувався періодами опору до впровадження штучного інтелекту в широку практику суспільної діяльності. Метою оригінального «тесту Тюрінга» було експериментально встановити, чи зможе людина розпізнати штучний інтелект, чи розуміє вона відмінності між людським мисленням та імітацією штучного інтелекту (A. Turing, 1950) [8]. Не зовсім вірною є інтерпретація цього тесту як спосіб підтвердження здатності машини «обманути людину» [7]. Такі завдання можуть ставитися лише під впливом дії антропоморфізмів та екстраполяції людських рис тим об'єктам, які їх мати не можуть. Питання, на які шукав відповіді Алан Тюрінг, були досить тривіальні: чи може людина відрізнити імітацію від оригінального людського мислення, що в подальшому визначатиме спосіб взаємодії людини з технологіями штучного інтелекту.

Виклад основного змісту статті. У 2000 році виходить стаття Turing Test: 50 Years Later, у якій описано повторне проведення «тесту Тюрінга». До цього часу технології штучного інтелекту далеко просунулися у своєму розвитку, а головне – отримали можливість доступу до велетенського масиву репрезентативних даних, яка дала змогу покращити генерацію відповідей на запити користувачів (Ayse Pinar Saygin, Ilyas Cicekli & Varol Akman, 2000). Штучний інтелект отримав здатність підлаштовуватися до індивідуальних і групових особливостей користувачів, і отримав можливість реагування на смислові підтексти, що, звісно, було неможливим в оригінальній версії тесту. Тренуючись на великих масивах даних, штучний інтелект також навчився всім соціальним упередженням і стигмам соціально сприйняття. Автором контенту, на якому тренується ШІ, є люди, тому він закономірно ідентифікує суттєві особливості відображення дійсності, не відчуваючи, що вони є порушеннями норм соціальної терпимості та толерантності (такі особливості сприйняття властиві тільки людині). Комп'ютер добре імітує людське мислення, якщо генерація відповідей відбувається на рівні простих когнітивних дефініцій: чим простіші алгоритми, тим менше відмінностей існує між відповідями людини та машини. Але людина переважно мислить розгорнуто, цей процес описується в когнітивній психології як «потік», який виникає внаслідок самодетермінації мисленнєвих процесів. Машина не здатна до такого, її робота із запитом має початок і завершення, а спроба побудови зациклених алгоритмів призводить до генерації беззмістовного абсурду. На момент опублікування дослідження його автори дійшли висновку, що в найближчому майбутньому штучний інтелект не зможе повною мірою імітувати людське мислення.

У 2016 році Джон Маккой і Томер Ульман із Массачусетського університету провели експеримент, на яких їх надихнув оригінальний «тест Тюрінга» [5]. Завдання досліджуваних (n = 1089) – довести одним словом, що вони не роботи, а люди. У кінцевому підсумку вони сформулювали мінімальний тест «людськість». За результатами асоціативного дослідження виявилося, що більшість асоціацій, які позначали суто людські атрибути, стосувалися сфери почуттів та емоцій: «любов» (134 особи), «співчуття» (33), «людськість» (30), «будь ласка» (25), «милосердя» (18), «співпереживання» (17), «емоція» (14), «робот» (13), «людство» (11), «живий» (9). Це свідчить про те, що штучний інтелект не може виражати емоції і досліджувані це усвідомлюють це повною мірою, четверта частина досліджуваних вибрала як показник «людськості» емоції, що спрямовані на іншого (інших). У другій частині експерименту досліджуваним запропонували проглянути пари слів, які виявили на першому етапі, та вибрати те, що вибрала людина. Найбільш повторюваним виявилося слово «кавелик» (зменшено-пестливе від «екскременти»), якому в жодному разі не приписали вибір штучного інтелекту. Це підтверджує те, що до генерації вульгарного абсурду може також вдатися людина, яка не обмежена рамками формальної логіки. «Любов», «кохання», «дружба», «ненависть» – це концепти, які дуже важко зрозуміти штучному інтелекту, навіть у процесі машинного навчання. Ці емоції сильно індивідуалізовані та слабо вербалізовані, вони не мають типових лексем і побудовані зі складних семантичних конструкцій.

У 2014 році у змаганнях на честь 60-річчя з дня смерті Алана Тюрінга вперше було пройдено його тест комп'ютерною програмою «Євген Густман» (серед розробників був українець Євгеній Демчак). Комп'ютерний співрозмовник переконав 33% журі змагань, що він реальний 13-річний хлопчик з Одеси (для проходження оригінального тесту потрібно 30%) [4]. Ідентифікація віку та місця проживання робилася не випадково, адже через вік співрозмовник міг формулювати прості відповіді й робити помилки в семантиці вживання слів.

Розвиток технологій «штучного інтелекту» відбувався хвилеподібно, основні підйоми та зниження зацікавленості суспільства були пов'язані з надмірними очікуваннями від них і розчаруванням, яке наступало після. Завдання, пов'язані з імітацією поведінки людини, з'явилися на пізніших етапах розвитку штучного інтелекту: до цього прямого завдання створити правдоподібну імітацію психічних процесів людини не було, вони виникали як результат екстраполяції властивостей живої матерії на неживу. Під час взаємодії з технічними системами людина почала ототожнювати себе з об'єктом своєї діяльності, наділяючи його не властивими йому рисами: емоціями, досвідом, цінностями, здатністю до розмірковування (здатністю мати свою думку). У такий спосіб відбулася спроба створити істоту, що може бути людиною, адже вона наділена «людськими рисами» з точки зору людини. **Технології не допомогли створити істоту, яка «думає та відчуває», це ми у своїх очікуваннях намагалися побачити в них те, чого там не було, видаючи, відверто кажучи, бажане за дійсне.**

За основними очікуваннями, які покладалися людством на розумні технічні системи, стояли сподівання на покращення людини як виду. Деякі з людських вад мали бути подолані завдяки розумному синтезу з технічними системами. У п'єсі Карла Чапека Rossumovi univerz ln roboti (перше видання 1921), від назви якої походить назва «робот», у комедійній формі

представлено суспільні уявлення про функціональні можливості (на той час футуристичні) розумних машин [2]. У ході розгортання сюжету показано, як бездушна технологія під кінець сюжету повстає проти людини, показуючи таким чином, що технічні системи можуть еволюціонувати до рівня людини та входити з нею у певне протиборство. За уявленнями майже сторічної давнини, роботи можуть навчатися, і це робить їх подібними до людини, адже ця подібність ілюструє еволюційний паралелізм формування людини та машини: підтвердження закону ілюстрації філогенезу в межах індивідуального онтогенезу. Розумні машини мали перевершити людину, таким чином проілюструвати, що еволюція не зупиняється на законах «дарвінізму», а може піти іншим шляхом – шляхом технологічного буму, який однозначно мав призвести до появи протиріч між розумними технічними системами та людиною.

Роботи або інші людиноподібні машини з психологічної точки зору мають одночасно антагоністичні риси: максимальну подібність до людини, але це є результатом дії антропоморфізмів як збоку користувачів, так і з боку конструкторів системи, і водночас відсутність позитивних, просоціальних рис, альтруїзму. Дослідницьке питання, чи взагалі є усвідомлення в пересічних користувачів ШІ розуміння того, що їх не можна порівнювати з людиною у будь-якому плані, особливо в психологічному, оскільки візуальна подібність не є запорукою наявності психологічної подібності. Босулов О. Ю. у статті «Потенційна небезпека штучного інтелекту», опираючись на матеріали публічних виступів засновників світових лідерів у використанні штучного інтелекту Microsoft, IBM, SpaceX, Apple, дійшов висновку, що непотрібно ставитися до ШІ як до предмета наукової фантастики, а його небезпека є об'єктивною [1]. З появою можливості автоматизації та креативності штучний інтелект матиме змогу самонавчатися, а це є основою еволюції. Проте яким чином відбуватиметься «репродукції» машин та чи зможуть вони обійтися без людини навіть як допоміжного виду, невідомо. Усе сходиться до того, що людство створює технічного монстра, який, набравшись певної самодостатності, починає загрожувати своєму творцю Тобто, **машина як уже самодостатній суб'єкт діяльності насамперед намагається знищити свого творця.** Ми не піддаємо сумніву, що технологічні винаходи останніх років становлять певну небезпеку, але чи є вони небезпечними як самостійний, не пов'язаний із людиною вид?

У нашому дослідженні ми дослідили специфіку атрибуції рис особистості стосовно штучного інтелекту, що з точки зору формальної логіки та наукової психології є виявом абсурду людського мислення, адже риси особистості можуть мати виключно люди, жодна інша жива (тварини) чи нежива (машини) істота їх мати не може. Але феномен людського мислення полягає в тому, що закони формальної логіки можуть провокувати системні порушення та спричиняти хибні закономірності. Досліджуваним пропонується десять рис особистості: п'ять позитивних – «доброта», «співпереживання», «чуйність», «працелюбність», «відповідальність» і п'ять негативних – «підступність», «жорстокість», «брехливість», «безвідповідальність», «інфантильність». Досліджувані мали оцінити, наскільки кожна з рис є властивою або для людини, або для машини (штучного інтелекту), використовуючи оціночну шкалу від 1 до 10, де 1 – риса, яка абсолютно властива штучному інтелекту, а 10 – риса, що властива виключно людині, інші проміжні варіанти (2–9) вказують на певну властивість риси ШІ та людині (більшою чи меншою мірою). Завдання дослідження полягали у вивченні впливу антропоморфізмів на сприйняття людиною штучного інтелекту. Риси особистості є суто людською властивістю та не можуть в об'єктивному плані бути екстрапольовані на технічні системи, отже, об'єктивною оцінкою кожної з рис особистості в межах нашого дослідження є 10, інші оцінки є результатом впливу антропоморфізмі. Наступним завданням дослідження є з'ясування того, якими рисами особистості наділяє людина машину, особливо якої модальності: позитивної чи негативної. Це важливо знати в плані розуміння того, як людина себе порівнює зі штучним інтелектом: вони є більш схожими чи протиставляються за рядом ознак.

У дослідженні взяли участь 87 осіб юнацького віку, які не взаємодіють зі штучним інтелектом у професійному плані, а отже, орієнтуються на стереотипні суспільні уявлення. У масовій культурі побутує певний образ штучного інтелекту, який відображає середньостатистичне ставлення до нього як суспільного феномену. Люди давно оцінюють машину як частину соціуму, яка займає певну частину суспільного простору та ставиться до неї не як до об'єкта, а наділяє рисами, які властиві людині. Це відбувається як результат каузального відображення причинно-наслідкових зв'язків. Людям хочеться, щоб те, що є результатом роботи програмного коду, мало і людське обґрунтування у вигляді пояснення притаманності машині властивостей, які здійснюють її саморегуляцію.

За результатами нашого дослідження ми не зафіксували рис, які є більш властивими штучному інтелекту, ніж людині (табл. 1). Але оскільки ми не отримали крайніх значень оцінки (10 балів, що свідчить про те, що риса властива тільки людині), це підтверджує нашу гіпотезу, що людині властивий антропоморфізм сто-

Таблиця 1

Результати оцінки властивості рис особистості людині та штучному інтелекту

Риси особистості	*Mo*	$\bar{X}$
Доброта	6	7,8
Співпереживання	5,7	6,7
Чуйність	9	9,5
Працелюбність	4,7	5,6
Відповідальність	6	6,2
Підступність	7,9	8,2
Жорстокість	6	6,4
Брехливість	7	5,3
Безвідповідальність	7	8,2
Інфантильність	8	8,3

совно технічних систем. Досліджувані найбільш близькими до штучного інтелекту вважають риси «працелюбність» і «брехливість», до найбільш суто людських рис належать «чуйність», «інфантильність» і «безвідповідальність».

Досліджувані не приписують штучному інтелекту негативних (антисоціальних) рис особистості, що також деякою мірою спростовує нашу гіпотезу про протиставлення людини та штучного інтелекту на рівні атрибуції рис особистості. Усі із запропонованих рис більш властиві людині, а приписування деяких рис штучному інтелекту опосередковано досвідом взаємодії з ним. З огляду на це можна зробити висновок, що досвід взаємодії людини із штучним інтелектом ґрунтується на професійних відносинах, де машина виконує певні завдання на високому рівні, перевершуючи можливості людини. У контексті цього можна також згадати першочергову мету створення штучного інтелекту: позбавлення людини від рутини, яка була малоосмисленим видом діяльності, що в більшості випадків демотивувало й у побутовому плані трактувалося широким загалом як «лінь». Тому найбільш відповідною рисою для пояснення роботи технічної системи є «працелюбність», яка має *Mo* = 4 та 7, що свідчить про те, що риса є більш властивою для технічних систем, ніж для людини (машина краще справляється з рутиною). Найбільш відповідною рисою для людини, на думку досліджуваних, є «чуйність» (*Mo* = 9) і «підступність» (*Mo* = 7 і 9). Ці риси стосуються сфери стосунків, які регулюються не процесуальними, а моральними імперативами. Ці риси особистості є властивими лише в системі людських взаємовідносин і не екстраполюються на систему каузальної атрибуції результатів діяльності технічних систем. Результати діяльності технічних систем (у тому числі штучного інтелекту) не можна пояснювати з точки зору дії суто людських рис особистості, які важко імітуються штучним інтелектом. Еволюція штучного інтелекту пішла в бік пошуку можливостей імітації діяльності людини: на початку появи технології – імітації людського мислення та на цей час імітації творчої діяльності, що є суто людським видом активності та не притаманне жодною мірою іншій живій істоті чи технічній системі [3]. Проте бажання поширити феноменологію людської поведінки на інші об'єкти зовнішнього світу призвело до появи віртуальних суб'єктів діяльності, які стали таким виключно через людське прагнення зрозуміти природу поведінки всього, що оточує людину (хоча з психологічної точки зору це не є актом свідомої саморегуляції).

Діяльність штучного інтелекту є імітацією психічної діяльності людини, не може з нею порівнюватися і, незважаючи на побідні результати діяльності, особливо інтелектуальної, отримується в різний спосіб. Особливістю людської психіки є здатність до генерації алгоритмів, які можуть містити логічні суперечності, які не будуть результатом помилки роботи алгоритму, а опиратимуться на інтуїтивні передбачення людини. У процесі генерації алгоритму діяльності людина, на відміну від штучного інтелекту, завжди допускає порушення (причому свідомі порушення), які призводять до появи нового й оригінального способу досягнення результату. Людині, якщо узагальнити попередні твердження, завжди кортить зробити «не так». Іншою, досить суттєвою, відмінністю штучного інтелекту та людської психіки є мотивація діяльності або діяльність як результат дії моральних імперативів. Саме цю властивість ми намагаємось екстраполювати на технології ШІ, щоб зрозуміти, яким чином машина дійшла до такого результату. Як приклад дії цього антропоморфізму людина під час взаємодії з технічними системами намагається в них «просити про допомогу», не піддаючи значенню те, що «допомагати» чи «не допомагати» є результатом дії моральних переконань людини.

Висновки. За результатами дослідження було встановлено, що під час оцінки роботи штучного інтелекту людина вдається до атрибуції не властивих йому особистісних рис із

метою кращого розуміння діяльності машини, особливо результатів діяльності. У досліджуваних не виявлено схильності до атрибуції особистісних рис певної модальності. Передбачалося, що досліджувані будуть приписувати негативні риси стосовно технологій штучного інтелекту, проте ця гіпотеза не підтвердилася. Дослідження дає змогу поглибити проблематику вивчення взаємодії особистості з технічними системами, особливо в умовах постійного зростання спектра можливостей, які вони розкривають для людства. Дослідження підтвердило гіпотезу про те, що технічні системи сприймаються нами як самостійні суб'єкти діяльності, які можуть продукувати поведінку, що ґрунтується на певних ціннісно-мотиваційних імперативах. Поява антропоморфізмів свідчить про тісну взаємодію людини з технічними системами як у професійному, так і в побутовому плані.

ЛІТЕРАТУРА:

1. Босулов О. Ю. Потенційна небезпека штучного інтелекту. *Інформація і право.* № 2 (14). 2015. С. 121–128.

2. Карел Чапек. Р. У. Р. (Россумові універсальні роботи). Київ : Видавничий дім «Комора», 2020. 192 с.

3. Кириченко В. В. Соціально-психологічна парадигма розуміння еволюції штучного інтелекту. *Психологія та соціальна робота.* 2023. № 2 (58). С. 17–25.

4. Eugene the Turing test-beating "human computer"–in"his"ownwords.*TheGuardian*Режимдоступу: https://www.theguardian.com/technology/2014/jun/09/eugene-person-human-computer-robot-chat-turing-test.

5. McCoy J., Ullman T. A Minimal Turing Test. *Journal of Experimental Social Psychology*. Vol. 79. 2018. P. 1–8.

6. Pinar Saygin A., Cicekli I., Akman V. Turing Test: 50 Years Later. *Minds and Machines.* 10, 2000. P. 463–518.

7. Teuscher C. Alan Turing: Life and Legacy of a Great Thinker. Springer; First Edition. 2003. 570 p.

8. Turing A. Computing Machinery and Intelligence. Mind, Vol. LIX, Issue 236. 1950. P. 433–460.