УДК 681.12

Д. О. Зелінська

ФІЛОСОФІЯ ЯК КЛЮЧ ДО СТВОРЕННЯ ШТУЧНОГО ІНТЕЛЕКТУ

Вінницький національний технічний університет

Анотація

Запропоновано обґрунтування тези, згідно з якою штучний інтелект не обов'язково повинен точно відтворювати устрій людської свідомості, щоб досягти високої ефективності й у перспективі позбавити людину ролі активної сили цивілізації. Вироблено набір обмежень, без застосування яких подальше вдосконалення ШІ виявиться згубним для людства.

Ключові слова: штучний інтелект, сильний штучний інтелект, тест Тюринґа, свідомість, етика.

Abstract

The substantiation of the thesis is proposed, according to which artificial intelligence does not necessarily have to precisely reproduce the structure of human consciousness, in order to achieve high efficiency and in the future to deprive man of the strength of active civilization. A set of limitations is developed, without which further enhancement of AI will prove to be disastrous for humanity.

Keywords: artificial intelligence, strong artificial intelligence, Turing's test, consciousness, ethics.

Безперечним є факт, що людський мозок має можливості в деяких відношеннях вищі, ніж у всіх інших відомих об'єктів у космосі. Досконалістю людського мозку обумовлена складність досягнення успіхів у розвитку загального штучного інтелекту, який був би аналогом мозку.

Сам термін «загальний штучний інтелект» є прикладом раціоналізації і узагальнення пошуків вчених у даній сфері, оскільки раніше область мала назву «штучний інтелект». Терміном «штучний інтелект» характеризували власне комп'ютерні програми, такі як ігрові програми, пошукові системи, системи розпізнавання, боти. «Загальний штучний інтелект» є набагато складнішим механізмом, що повинен бути максимально наближеним за ознаками до людського розуму.

Попри те, що вченим поки не вдалося створити загальний штучний інтелект, ця задача може бути розв'язана. Закон про універсальність обчислень говорить, що будь-який процес, заснований на фізичних законах, може бути відтворений програмою відповідної складності, якщо забезпечити їй достатньо часу і енергії. Те, що штучний інтелект має бути "людиною", обговорювалось в його концепції з початку. Якщо програмі буде недостатня хоч якась когнітивна здатність людини, вона не потрапить під визначення «загального штучного інтелекту», а використання некогнітивних ознак для визначення людськості (наприклад, процентний вміст вуглецю) стане расизмом. При цьому, не можна боятися визначати об'єктивні відмінності між людьми та іншими мислячими істотами - ці відмінності повинні грати життєво важливу роль у цивілізації, що включає носіїв «загального штучного інтелекту».

Девід Дойч, професор фізики Оксфордського університету, вважає що всі поширені точки зору на штучний інтелект містять ряд фундаментальних помилок. По-перше, принижено цінність штучного інтелекту - деякі вважають, що він навряд буде "розумнішим", ніж вже існуючий софт. Крім того, люди марнославні і хочуть залишатися найближчими до ідеальних істот. По-третє, занадто велике значення в проблемі ЗШІ надається самосвідомості та свідомості - при тому, що свідомість попередньо залишається дуже нечітко визначеним терміном.

З попередніх тез можна зробити висновок, що проблема штучного інтелекту - це проблема філософії, а не комп'ютерних наук або нейрофізіології, і філософський прогрес зіграє велику роль у вирішенні цього завдання. З погляду Дойча, ЗШІ – творча істота, яку неможливо створити без розуміння якісних відмінностей між власне ним та звичайною комп'ютерною програмою. [1]

Очевидно, що перш ніж говорити про ЗШІ, потрібно розібратись з поняттям людського розуму. Інтуїтивний підхід стверджує, що в певному сенсі існують два аспекти існування людини – фізичний і нефізичний. Фізичний вимір - це людське тіло - не тільки ноги та руки, а й мозок. Робимо припущення,

що мозок не впливає на наші психічні стани та мислення. Наш нефізичний чи нематеріальний вимір - це місце, де відбуваються ці психічні стани, мислення, емоції, тобто розум. Звідки випливає, що розум і тіло існують окремо одне від одного. Попри розділеність, розум і тіло можуть взаємодіяти, і це поняття отримало назву дуалізму. Рене Декарт, один з найвідоміших філософів XVII століття, створив власну теорію дуалізму - так званий "картезіанський дуалізм". Декартів дуалізм, як і звичайний дуалізм, стверджує, що розум і тіло є окремими субстанціями, що мають взаємно причинні зв'язки. Однак Декарт стверджує, що душа існує в окремій площині (тобто не тій самій площині чи реальності, в якій існує світ), але на неї якимось чином може впливати тіло (через мозок), а сама душа також може впливати на тіло (теорія союзу відмінних одна від одної субстанцій). Проте, найслабшим місцем картезіанського дуалізму є неспроможність довести суттєву відмінність розумових можливостей людини від розумових можливостей інших об'єктів всесвіту. Вчення Декарта рятує людську свободу, але не може пояснити точних механізмів взаємодії душі й тіла (він описує цю взаємодію, але не пояснює, чому саме вона може відбуватися). [2]

Саме ці сумніви привели відомого британського вченого Алана Тюринґа до питання «чи можуть машини робити, що ми (як мислячі істоти) можемо робити?». За Тюринґом, математичні межі логіки і обчислень можуть істотно обмежити інтелект обчислювальних машин. Він стверджує, що «є ряд результатів математичної логіки, які можна використовувати, щоб показати, що повноваження дискретних машин істотно обмежені». Математична межа пов'язана з іншими обмеженнями розумних машин і є «аргументом на користь свідомості». Тюринґ пропонує дослідити гру з двома учасниками, один з яких є людиною, а інший - програмою. Ідея тесту Тюринґа полягає в тому, що людина «всліпу» має визначити, хто є супротивником: програма чи людина. Якщо людина-учасник тесту не відрізнить людину від програми, то остання і буде штучним інтелектом.

Проте, постає питання чи можна вважати людський мозок комп'ютерною програмою. Американський філософ Джон Серль критикує тест Тюринґа за допомогою уявного експерименту «Китайська кімната». Серль порівнює запитання тесту Тюринґа з китайською мовою і ставить себе на місце програми, що повинна пройти тест на розуміння. Навіть не знаючи китайської мови, учасник уявного експерименту буде спроможним «відповісти» на запитання і «витримати» тест на розуміння мови. Але ж це не означає, що учасник насправді володіє мовою, а програма Тюринґа справді є ШІ.

При розробці експериментів Серль увів у науку поняття «сильний штучний інтелект» як абсолютно новий термін. Саме такий інтелект зможе виконати будь-яку інтелектуальну задачу, яку може виконати людина. За словами Серля, «така програма буде не тільки моделлю розуму; вона в буквальному розумінні слова сама і буде розумом, в тому ж розумінні, в якому людський розум — це розум..».[3, 107]

Зворотньою стороною медалі створення «сильного штучного інтелекту» або ж «загального штучного інтелекту» є низка проблем з морально-етичним аспектом. Наприклад, якщо штучний інтелект - програма, що працює на комп'ютері, видалити її з комп'ютера - вбивство, так само, як і позбавити людський розум фізичного тіла. Але штучний інтелект може бути скопійований на безліч комп'ютерів одним натисканням кнопки. Чи будуть ці програми, запущені на різних комп'ютерах, однією і тією ж особою або різними людьми? Як враховувати їхні голоси на виборах? Якщо говорити про представників штучного інтелекту як про творчих істот, то з ними не можна поводитися як з іншими комп'ютерними програмами - це означало б «промивання мізків» і тиранію. Ігнорування прав і індивідуальності ШІ, буде не тільки злочином, а й джерелом проблем: творчі істоти не можуть вічно існувати в рабстві.

Ще одним аргументом на користь неможливості створення ЗШІ є те, що за релігійним вченням, мислення є функцією безсмертної душі людини. Бог дав безсмертну душу кожній людині, але не програмі чи механізму. Тому жодна програма не може мислити. Звідки випливає, що навіть якщо машина відтворить всі внутрішні механізми мозку розумної істоти, вона не зможе досягти справжнього інтелекту без душі.

З огляду на різні точки зору і аспекти, виникає питання чи справді сильний штучний інтелект повинен мати всі властивості людської свідомості й бути повним її аналогом. Адже, надання ШІ абсолютно усіх людських якостей матиме як переваги, так і недоліки. Так, надання ШІ таких якостей, як емпатія, почуття гумору, людяність, здатність до новаторства було б великим кроком вперед у плані взаємодії з людством. Але водночас, разом з цими позитивними «людськими» якостями з'явились би і роботи з цілком протилежними рисами характеру, а це становитиме значну загрозу. Надання їм повної свободи може означати, що ЗШІ захопить владу на Землі, що може закінчитись трагічно для їх же творців. Проте ця проблема має мало спільного зі штучним інтелектом. Боротьба між ідеями добра і зла існує вічно і не залежить від фізичного «обладнання», на якому вона протікає. Суть в тому, що ми хочемо,

щоб «добрий розум», в будь-якому його вигляді, перемагав «злий розум», але наша концепція добра потребує постійного поліпшення. «Поневолити всіх розумних істот» - катастрофічно неправильне вирішення проблеми, а «поневолити всіх розумних істот, не схожих на нас» звучить не набагато краще. Проте ШІ повинен мати відповідний кодекс, як і людина має слідувати певним заповідям і нормам.

Важливо врахувати і те, що людський інтелект має дуже складну будову. Дослідники з Массачусетського технологічного інституту близько 16 років тому висунули гіпотезу, що нейронні мережі є основою психічних здібностей. Так виявилось, що у мозку людини є великий обсяг окремих інструкцій, що координуються разом. Без нейронних мереж сумнівно чи взагалі існували б процеси мислення, мовні здатності та свідомість. Попри значний успіх досліджень, вчені досі не мають повного розуміння роботи та будови нейронних мереж, і тому обчислювальні системи створюють на основі примітивніших нейронних мереж тварин. Крім того, штучні нейронні мережі є крихітними і простими порівняно з біологічними аналогами: десятки тисяч нейронів у порівнянні з трильйоном.

Наскільки штучний інтелект має бути схожим на людський – на це питання ми отримали відповідь вище з точки зору нейрофізіології. Надзвичайна складність створення аналогу людського мозку говорить про те, що відмінності будуть, і досить суттєві. Однак справа не у відмінностях, а в результатах функціонування: хай яким би був ШІ, його активність має визначатися певними обмеженнями (які будуть своєрідним аналогом людських етичних правил):

1. ШІ має бути безпечним для себе і для людей, а при виникненні помилок у роботі - надавати змогу усунути їх наслідки.

2. ШІ не повинен обмежувати свободу волі і приватність будь-кого. Хоч ми і говоримо про можливість розвитку ШІ як самостійної творчої істоти, потрібно проектувати дані системи з установкою сумісності людських ідеалів гідності, прав, свобод, та, що є дуже важливим, толерантності. Толерантність має стати основною рисою, якою оволодіють усі роботи без виключення, адже саме цієї риси не вистачає часто навіть у прогресивному гуманістичному суспільстві.

3. ШІ не можна використовувати для підривної діяльності й гонки озброєнь. Влада, що отримується за допомогою контролю високорозвинених систем ШІ, повинна поважати і покращувати, а не підривати соціальні і громадянські процеси, від яких залежить здоров'я суспільства.

Отже, штучний інтелект повинен розроблятись на благо всього людства та задля формування такого рівня сучасного суспільства, якого люди ще не досягли самостійно. На наш погляд, більшість людства зацікавлена саме в такому розвитку ШІ, але цю програму буде не легко втілити, оскільки ШІ вже сьогодні використовується для різноманітних спецоперацій та у військових технологіях. Можливо, сутність людської свідомості ніколи не буде пізнана. Але людству, якщо воно не хоче занапастити себе небезпечним використанням ШІ, доведеться виробити такі механізми співпраці й такі філософські моделі майбутнього, які усунуть можливі загрози від ШІ.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1.Philosophy will be the key that unlocks artificial intelligence // The Gurdian Post, October 3, 2012

2.URL:https://ayearofai.com/rohan-7-can-artificial-intelligence-achieve-human-intelligence-b0c95e23ca4b

3.Сепетий Д.П. - Свідомість як суб'єктивність: таємниця Я. – 2-ге вид., перероб. і доп. – Книга 1. Зомбі, комп'ютери та Абсолютний Дух. – Запоріжжя: Просвіта, 2017. – 304с.

Зелінська Дарія Олегівна **–** студент групи 2КН-16б, Факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: dariia050500@gmail.com

Науковий керівник: ***Хома Олег Ігорович*** — д-р філос. наук, професор, завідувач кафедри філософії та гуманітарних наук, Вінницький національний технічний університет, м. Вінниця

Zelinska Dariia — student of the 2KN-16b group, Faculty of Information Technologies and Computer Engineering, Vinnytsa National Technical University, Vinnytsa, e-mail: dariia050500@gmail.com.

Supervisor: ***Khoma Oleg I.*** — Dr. Sc. (Philos.), Professor, Head of the Chair of Philosophy and Humanities, Vinnytsia National Technical University, Vinnytsia