УДК 1/004.8

В. О. Кривенька

Що таке «сильний штучний інтелект»? Аргумент Джона Серла проти тесту Тюринга

Вінницький національний технічний університет

Анотація
У статті розглянуто суть тесту Тюринга, а також аргумент проти цього тесту, запропонований Джоном Серлом у формі мисленнєвого експерименту «Китайська кімната». Я доводжу, що цей аргумент поки що переконливо спростовує можливість сильного штучного інтелекту. Але ми не розуміємо, що таке людська свідомість. Отже, ми не можемо впевнено сказати, що в принципі не можуть бути створені системи, аналогічні «свідомості».
Ключові слова: штучний інтелект, робот, комп'ютер, свідомість, тест Тюринга, феноменальні зомбі.

Abstract
The paper analyzes the nature of the Turing test and the argument against this test proposed by John Searle in the form of a «Chinese Room» thought experiment. I argue that this argument so far convincingly denies the possibility of strong artificial intelligence. But we do not understand what human consciousness is. Therefore, we cannot say with certainty that, in principle, the systems similar to "consciousness" cannot be created.
Keywords: artificial intelligence, robot, computer, consciousness, Turing's test phenomenal zombies.

Власна логіка дослідницького процесу в галузі штучного інтелекту (далі – ШІ) призвела до питання, чи можливо створити ШІ, який принаймні було би важко відрізнити від людини, а в ідеалі – який став би істотою, що мислить, хоч і не має людського мозку. На відміну від різноманітних ужиткових версій ШІ, цей отримав назву «сильного» ШІ. Отже, ідеться про принципово різні завдання: створити щось, який розв'язує якісь окремі завдання (тобто – створити інструмент) і щось, що саме ставить завдання (тобто – творця). Цей так званий робот має вміти поводитись як людина, реагувати, спілкуватись, обробляти інформацію та навчатись. Але, до того ж, усе це він має робити «як ми», люди (тобто, не імітувати нас, а «справді» розуміти все, що він робить, придумує тощо), хоч і на іншій, так би мовити, «елементній базі».

Саме так постала проблема свідомості ШІ, адже саме «свідомість» вважається тим, що відрізняє людину він тварин чи автоматів. Я не зупинятимуся на питанні відсутності однозначного визначення свідомості серед філософів і вчених, оскільки це не є безпосередньою складовою моєї теми. Зазначу лише, що як літаки можуть літати, не маючи при цьому ані пташиного пір'я, ані крил, якими можна махати, так і, можливо, ефекти «свідомості» можуть бути відтворені без того, що властиве людському «носієві» (тіло, мозок, соціалізація, культура тощо). Чи можна відтворити реальні ефекти свідомого мислення, не розуміючи, чим воно є? Цим питанням, зокрема, займався англійський математик Алан Тюринг. У 1950 році він створив славнозвісний тест, названий пізніше його ім'ям [4]. Щоб не потерпати від майже нерозв'язного питання визначень: що таке інтелект (свідомість, мислення тощо)? Де межа між машинною обробкою інформації і мисленням? Тощо, Тюринг вирішив сформулювати питання інакше, у чіткіших межах [9, p. 433]. Передбачалося створити процедуру визначення того, наскільки машина здатна виявляти інтелектуальну поведінку, тотожну людській, тобто таку, що її неможливо відрізнити від людської.

Я коротко опишу суть цього тесту, орієнтуючись на стандартне формулювання, запропоноване у ювілейній статті з журналу «Minds and Machines» [6]: Людина, що грає роль оцінювача, спілкується з одним комп'ютером і з однією людиною, не знаючи, хто є його співрозмовником (ідеться не про голосову розмову, а про обмін повідомленнями). На підставі відповідей на питання оцінювач повинен визначити, із ким він спілкується: з людиною чи з комп'ютерною програмою? Завдання комп'ютерної програми – не дати оцінювачу приводу визначити, що він спілкується не з людиною, тобто – не дати розкрити себе.

Комп'ютер успішно пройде тест Тюринга, якщо людина-експериментатор, поставивши йому письмово певні питання, не зможе визначити, від кого отримані відповіді: від іншої людини чи від комп'ютера. Складання програми для проходження тесту Тюринга вимагає великого обсягу роботи. Насамперед комп'ютер повинен мати [4]: засоби обробки текстів, висловлених природними мовами, наприклад англійською; засоби автоматичного формування логічних висновків, що забезпечують можливість використовувати збережену інформацію для пошуку відповідей на питання і отримання нових висновків; засоби машинного навчання, які дозволяють пристосовуватися до нових обставин, тощо.

25 жовтня 2017 року Саудівська Аравія стала першою країною, що надала громадянські права роботу. Робот Софія стала головним творінням Девіда Хенсона, відомому завдяки виготовленню людиноподібних роботів. Софія може імітувати людські вирази обличчя, відповідати на певні запитання, вчитися та вдосконалюватися. Отож, постає питання чи можна уподібнити свідомість комп'ютерній програмі? Якщо робот поводиться як людина, це ще не означає що він має свідомість та вміє мислити. Штучний інтелект працює так, як його було запрограмовано людиною. Дехто вважав, що тест Тюринга змогла пройти програма Еліза (1966), але в цьому випадку просто йшлося про першу програму, яка підтримує ілюзію людського спілкування, доволі примітивно, до речі. Принаймні ані золоту, ані срібну медалі Лобнера для програм, що насправді пройдуть тест Тюринга, поки не вручено. Принцип роботи Елізи полягав у дослідженні фраз співрозмовника на наявність ключових слів. Застосовуючи свій алгоритм, програма могла вводити в оману деяких людей, що думали, ніби вони розмовляють із реальною людиною. Проте це не означає, що відповіді Елізи є свідомими. Останніми роками з'являлося не одне повідомлення про проходження тесту. Але тут ми маємо справу лише з різними спробами введення оцінювачів в оману, а не з мисленням. Попри сімдесятирічний термін існування тест Тюринга не втратив своєї значущості. Але це значення поступово стає більше філософським, ніж технологічним, оскільки дослідники воліють докладати зусилля не до імітації людських відповідей (т.зв. «штучна тупість»), а до дослідження реальних властивостей і проявів того, що ми називаємо інтелектом.

Усіх цікавить головним чином можливість сильного ШІ. На відміну від «слабкого ШІ», «сильний» буде вміти самостійно навчатися. Ідеться про інтелект машини, який може успішно розв'язати будь-яку інтелектуальну задачу, що її може розв'язати людина. Термін «Сильний ШІ» увів Джон Серл 1980 року: «…комп'ютер – не просто засіб дослідження розуму (mind), радше сам належним чином запрограмований комп'ютер насправді є розумом, у тому сенсі, що про комп'ютери, які наділені коректними програмами, можна буквально сказати, що вони розуміють і мають інші когнітивні стани» [7, p. 417]. Серл запровадив цей термін для позиції, яку піддав критиці.

На сьогодні питання про свідомість комп'ютера так і залишається відкритим. Деякі стверджують, що свідомість властива лише людям, аж ніяк не машинам. Цієї ж позиції дотримувався й Серл. Для підтвердження своїх поглядів він запропонував уявний експеримент «Китайська кімната», покликаний прямо заперечити тест Алана Тюринга.

Суть експерименту: людина, що не знає китайської мови, потрапляє в ізольоване приміщення разом із книгою, де вказані точні інструкції щодо складання ієрогліфів, але без пояснення їх сенсу. Той, хто знає китайську, знаходиться за дверима і через щілину під дверима передає на аркуші питання, написані китайською. Сама інструкція, розміщена в книзі, містить не пояснення того, про що йдеться у листі, а покрокову інструкцію складання відповіді. Ця інструкція є алгоритмом, подібним до комп'ютерного. Отримавши формально «правильну» відповідь, людина по той бік дверей може вважати, що людина всередині надала усвідомлену відповідь на питання, утім, Серл вказує на те, що людина від механічного компонування знаків не стала розуміти китайську, отже – не усвідомила, про що її питали й що вона відповіла [7, p. 421; див. опис у 1].

Уявний експеримент Серла ясно показує, що існує принципова відмінність між справжнім розумінням і його імітацією – навіть якщо це досконала імітація (симуляція), яку (ззовні) неможливо відрізнити від оригіналу. Комп'ютерні програми не мають нічого спільного з розумінням [див. 3, с. 106-107]. За Серлом, комп'ютери виконують лише формальні маніпуляції з символами; ці маніпуляції навіть важко назвати маніпуляціями з символами, бо символи нічого не означають для самого комп'ютера [7, p. 422]. Отже, вміння маніпулювати символами замало, щоби гарантувати знання, сприйняття. За Серлом, існування самої лише комп'ютерної програми недостатньо, щоби можна було говорити про наявність знання [див. 1].

Люди прагнуть створити ШІ, який буде імітувати поведінку людини. Тобто він оброблятиме той самий тип інформації, реагуватиме «людським» чином на вхідні дані, а його поведінка не відрізнятиметься від людської. Нам буде здаватися, що він є тотожний людині. Проте це не зовсім так. Річ у тому, що він цілковито позбавлений феноменальних відчуттів. Він, за Девідом Чалмерсом, зовні схожий на нас, усередині містить лише суцільну темряву [див. 5, с. 62; див. також 2, с. 64].

Чалмерс надзвичайно популяризував проблему «феноменальних зомбі» [5, с. 61]. Це такі істоти, яких неможливо відрізнити від нормальної людини. Проте вони відрізняються від людей тим, що не мають свідомості, досвіду або здатності щось відчувати. Наприклад якщо зомбі, коле себе гострим предметом, він не відчуває болю. У той же час він поводиться так, ніби відчуває його (він може сказати «ай» і відскочити або сказати, що йому «боляче»), хоча в нього, звісно ж, немає сприйняття болю, як у людини.

Аргумент зомбі є схожим на експеримент із китайською кімнатою. Інколи ми можемо навіть не здогадуватися, із ким спілкуємося. Нам будуть давати чіткі, логічні, змістовні відповіді, проте чи є вони усвідомленими?

Чим швидше з'являються нові ШІ, тим більше непокоїтимуться люди, побоюючись широко представленого у фантастичній літературі «повстання роботів». Дехто вважає, що одного дня всі машини можуть «повстати» і встановити свою «владу» над людством. Але ж для цього їм потрібно бути розумнішими за людей. Ситуація тут зовсім не «літературна» й не така проста, як здається. Як казала найбагатша людина Китаю, успішний бізнесмен Джек Ма: «Ви дійсно думаєте, що людство зможе створити щось таке, що буде розумнішим за нього? У машин є чіпи, а в людей – серця. Ніщо не замінить людського серця. У майбутньому машини стануть розумнішими, вони не втомлюватимуться та краще навчатимуться. Проте їх не потрібно боятися, адже ми завжди зможемо стати кращими за них» [5а].

На мою думку, обґрунтованим є висновок, що ШІ можна навчити імітувати поведінку людини, обробляти дані, вчитися, спілкуватися, але неможливо навчити його *усвідомлювати* зроблені ним дії. ШІ не перебував у процесі природної еволюції, і це, за М. Сіпером і Д. Муром, є його вразливим місцем ШІ. У своєму дописі в Ньюсвік вони посилаються Джуду Перла, лауреата премії Тюринга, за словами якого ШІ добре схоплює закономірності, на кшталт «Люди, що купують зубну пасту, купують і зубні щітки». Однак йому важко відповісти, чому люди купують зубні щітки або ж – на питання «якщо клієнт не купує пасту, чи купить він зубну щітку?» Тому ШІ поки зарано мріяти про захоплення світу чи принаймні про повну довіру від людей, які не можуть на нього в усьому покластися. Перш ніж панувати над світом, варто здогадатися, навіщо люди взагалі купують зубну щітку і пасту [див. 8]. Отже, можна впевнено сказати, що роботи є лише машинами, які діють за певним алгоритмом, обробляючи дані.

Однак можна поставити питання гостріше: що буде, якщо процес еволюції теж можна буде якось змоделювати (чи створити якийсь його аналог) і долучити до «виховання» нових типів ШІ? Якого прогресу нам чекати в цій галузі? На мій погляд, наукова нез'ясованість проблеми поки не дає можливості для переконливої відповіді. Нині більш імовірним постає варіант неможливості створення сильного ШІ. Утім, для сучасних людей актуальніше думати не про машин, які думатимуть «як люди», а про себе самих – чи завжди ми думаємо «як люди»? Це дуже актуальна проблема, адже різні стереотипи, мисленнєві звички тощо поглинають переважну більшість нашого часу, тож це велике питання – наскільки часто ми самі виявляємося істотами, що «мислять»?

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Клокун А. Джон Серль: штучний інтелект і «розумність». URL: https://plomin.club/searle-john-artificial-intelligence-and-reasonableness/
2. Леонов А. «Аргумент зомбі» Девіда Чалмерса: переднє слово перекладача Філософська думка, 2015 № 5, С. 51–59.
3. Сепетий Д. П. Свідомість як суб'єктивність: таємниця Я. – Книга 1. Зомбі, комп'ютери та Абсолютний Дух. – Запоріжжя: Просвіта, 2017.
4. Таранська М. Тест Тюрінга. URL: https://vido.com.ua/article/13520/tiest-tiuringha/
5. Чалмерс Д. Аргумент 1: Логічна можливість зомбі, 2015 № 5, С. 60–67.
5а. Чи зможе штучний інтелект перевершити людину: Джек Ма про майбутнє. URL: https://24tv.ua/techno/chi_zmozhe_shtuchniy_intelekt_pereversh iti_lyudinu_dzhek_ma_pro_maybutnye_n1231107
6. Saygin, A. P.; Cicekli, I.; Akman, V. Turing Test: 50 Years Later" (PDF), *Minds and Machines*, 2000, **10**(4): 463–518

7. Searle J. Minds, Brains, and Programs. Behavioral and Brain Sciences, 1980, Vol. III, Iss. 3, P. 417–424. https://doi.org/10.1017/S0140525X00005756
8. Sipper M., Moore J. H. The aliens have landed – but they're not smart enough to take over. Newsweek, 24.1.2019. URL: https://www.newsweek.com/aliens-invasion-artificial-intelligence-ai-takeover-evolution-1303841
9. Turing A. Computing Machinery and Intelligence. Mind, 1950, № LIX (236), P. 433–460. https://doi:10.1093/mind/LIX.236.433

Кривенька Вікторія Олегівна - студент групи 2КІ-18б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: vicka0701@ukr.net

Науковий керівник: Хома Олег Ігорович — д-р філос. наук, професор, завідувач кафедри філософії та гуманітарних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: sententiae2000@gmail.com

Kryvenka V. O. — student of the 2KI-18b group, Faculty of Information Technologies and Computer Engineering, Vinnytsa National Technical University, Vinnytsa, e-mail: vicka0701@ukr.net

Supervisor: Khoma O. I. — Dr. Sc. (Philos.), Professor, Head of the Chair of Philosophy and Humanities, Vinnytsia National Technical University, Vinnytsia, e-mail: sententiae2000@gmail.com