https://doi.org/10.15407/fd2026.01.103
УДК 130.1

Руслан МИРОНЕНКО,
магістр філософії,
засновник освітнього проєкту «Печера Платона»,
03057, Київ, вул. Євгенії Мірошниченко, 6/11;
аспірант кафедри логіки, філософський факультет,
Київський національний університет імені Тараса Шевченка,
01601, Київ, вул. Володимирська, 64/13;
провідний інженер відділу соціальної філософії
Інституту філософії імені Г.С. Сковороди НАН України,
01001, Київ, вул. Трьохсвятительська, 4
mironrus@gmail.com
https://orcid.org/0000-0003-4058-9772
SCOPUS ID: 57206845047

АРХІТЕКТОНІКА «КИТАЙСЬКОЇ КІМНАТИ»: РЕКОНСТРУКЦІЯ ТА ОЦІНКА МИСЛЕННЄВОГО ЕКСПЕРИМЕНТУ ЗАСОБАМИ ТЕОРІЇ АРГУМЕНТАЦІЇ

У статті проаналізовано відомий мисленнєвий експеримент Джона Серля «Китайська кімната», який заперечує здатність штучного інтелекту до справжнього розуміння. Дослідження має на меті крок за кроком розібрати цю аргументацію, щоб з'ясувати, як саме вона побудована, у чому її переконливість та де криються слабкі місця. Для цього застосовано модель, яка розглядає текст на чотирьох рівнях. На першому рівні з'ясовано історичне тло: проти чиїх саме ідей та комп'ютерних програм виступав автор. На другому рівні показано, як Серль вибудовує сценарій експе-

Цитування: Мироненко, Р. (2026). Архітектоніка «Китайської кімнати»: реконструкція та оцінка мисленнєвого експерименту засобами теорії аргументаціїю. *Філософська думка*, 1, 103—125. https://doi.org/10.15407/fd2026.01.103

рименту, щоб непомітно запропонувати авдиторії вигідні для себе правила дискусії. На третьому етапі перевірено релевантність порівняння роботи комп'ютера з діями людини, яка механічно сортує незрозумілі їй ієрогліфи. Аналіз виявляє, що переконливість аргументації значною мірою тримається на припущенні: нібито лише біологічний мозок здатний породжувати свідомість. На четвертому рівні експеримент перевірено на стійкість за допомоги найвідоміших заперечень критиків. Здійснений аналіз засвідчує, що Серль успішно доводить нездатність машини розуміти сенс лише через механічне переставлення символів. Проте його мисленнєвий експеримент виявляється вразливим до припущення, що свідомість може виникати як абсолютно нова властивість надскладних систем. На прикладі сучасних великих мовних моделей (LLM) зроблено висновок: аргумент «Китайської кімнати» досі актуальний, оскільки доводить, що машини не усвідомлюють фізичного зв'язку слів із реальним світом. Водночас це не виключає того, що штучний інтелект здатний успішно оперувати синтаксичними зв'язками між словами всередині самої мови.

Ключові слова*: «Китайська кімната», Джон Серль, теорія аргументації, філософія свідомості, неформальна логіка, риторика, мисленнєвий експеримент, міркування за аналогією, комп'ютаційний функціоналізм, метааргументація, прагма-діалектика, сильний ШІ, інференційна семантика, LLM.*

Вступ

Від часу першої публікації у програмній статті Джона Серля «Розум, мозок і програми» (Searle, 1980) мисленнєвий експеримент «Китайська кімната» остаточно набув статусу одного з найбільш обговорюваних у філософії свідомості та когнітивній науці. Як засвідчують класичні оглядові праці, наприклад, антологія за редакцією Дж. Престона та М. Бішопа (Preston & Bishop, 2002), навколо цього аргументу сформувалася безпрецедентна за обсягом академічна бібліографія. Утім, незважаючи на масштабність полеміки, переважна більшість дослідників традиційно підходила до аналізу «Китайської кімнати» крізь доволі вузьку оптику, зосереджуючись майже виключно на формальній перевірці правильності мисленнєвого експерименту або на спробах прямого спростування його висновків.

У межах цього мейнстрімного «логіко-епістемологічного» підходу мисленнєвий експеримент зазвичай редукується до набору пропозиційних засновків із метою виявлення логічних хиб або семантичних неточностей. Показовою у цьому контексті є праця Б.Дж. Коупленда, який розглядає експеримент суто з «логічного погляду» та намагається діагностувати в ньому помилку двозначності (fallacy of equivocation) (Copeland, 2002). Ларі Гаузер у своїх дослідженнях (у розвідці "Nixin' Goes to China") ставить за мету розвінчати (debunk) аргумент Серля шляхом прискіпливого аналізу семантичних понять (Hauser, 1997). Схожою траєкторією рухається Дейл Жакет, пропонуючи строгий формальний аналіз проблеми інтенційності у своїх «Пригодах у Китайській кімнаті» (Jacquette, 1989). Окрему лінію критики репрезентують Пол та Патриція Черчленди, які атакують висновки Серля через конструювання контраналогії зі «світною

кімнатою» (luminous room) та реконструюють його міркування як циклічні (Churchland & Churchland, 1990). До цієї ж традиції належить і критика Джері Фодора з позицій гіпотези «мови мислення» (language of thought) та природи обчислювального синтаксису (Fodor, 1987, 1991).

Суттєва прогалина таких підходів полягає в тому, що надмірно концентруючись на істинності окремих засновків (насамперед на аксіомі про те, що синтаксису принципово недостатньо для семантики), вони ігнорують аргументативну динаміку та комунікативний контекст експерименту. У таких статичних реконструкціях «Китайська кімната» постає як застигле міркування, штучно відірваний від реальної діалектичної ситуації живої академічної суперечки.

У цій статті я поставив на меті заповнити цю методологічну прогалину, протиставивши традиційному статичному аналізу принципово новий підхід. Головна теза дослідження полягає в тому, що «Китайську кімнату» слід розглядати не як ізольовану логічну задачу, а як складну аргументаційну архітектоніку. Для її реконструкції та оцінки у статті застосовано чотирирівневу аргументаційну модель, що розгортається за каскадним принципом: від зовнішнього макроконтексту дебатів до мікроструктури внутрішніх міркувань та їхньої діалектичної стійкості.

На першому рівні залучається оптика метааргументації Мориса Фінок'яро (Finocchiaro, 2013). Це дає змогу осягнути історичну макроструктуру дебатів, ідентифікувати реальну мішень Серля та розглянути експеримент як комунікативну ситуацію, де автор виправдовує власні інференції через критику епістемічних зобов'язань опонентів. На другому рівні, спираючись на прагма-діалектичний підхід Євгена Попи (Popa, 2016), мисленнєвий експеримент постає вже не як набір статичних засновків (наприклад, як редукований modus tollens), а як динамічна подія. Сценарій «Кімнати» розкривається тут як акт стратегічного маневрування на стадії «відкриття» дискусії, що диктує авдиторії специфічні правила гри та інтерпретації.

Третій етап присвячено перевірці внутрішньої правильності аргументації за допомоги структурної теорії аналогії Пола Барти (Bartha, 2010). Цей інструментарій уможливлює діагностику причинно-наслідкових зв'язків між джерельним та цільовим доменами експерименту, оприявнюючи залежність міркувань від прихованої онтологічної передумови біологічного натуралізму. Нарешті, на четвертому рівні здійснюється формалізація конструкції Серля за допомоги схем аргументації та критичних запитань Дуґласа Волтона (Walton, Reed, & Macagno, 2008). Цей підхід слугує своєрідним структурованим діалектичним стрес-тестом, який дозволяє оцінити стійкість захисту експерименту проти ключових контратак функціоналістів (зокрема «Системної відповіді» та «Відповіді робота»).

Зміна оптики з класичної логіко-епістемологічної на запропоновану багатовимірну аргументативну модель дає змогу уникнути методологічного еклектизму: результати реконструкції одного етапу формують інтерпретативну рамку для наступного. Такий підхід виводить дослідження за межі простого пошуку формальних хиб і розкриває справжні механізми переконливості, концептуальні пастки та діалектичну життєздатність експерименту Джона Серля.

Рівень I. Метааргументація

Оптика історико-текстуального підходу (М. Фінок'яро) спонукає трактувати цей мисленнєвий експеримент як взірець метааргументації. Згідно з визначенням Фінок'яро, метааргументація (meta-argument) — це «аргументація про один або більше аргументів», на відміну від аргументації базового рівня (ground-level arguments), які стосуються об'єктів чи явищ реального світу (Finocchiaro, 2013: p. 1). У випадку з мисленнєвим експериментом, об'єктом аналізу стає не фізична реальність (кремнієві чіпи чи нейрони), а структура аргументації опонентів. Цей метод вимагає відмовитися від аналізу «штучних» прикладів на користь детального вивчення реальних теоретичних суперечок у їхньому автентичному контексті. Як зазначає Фінок'яро, такий аналіз передбачає не лише інтерпретацію, але й «обґрунтовану оцінку аргументації» (reasoned evaluation of arguments), що перетворює сам процес дослідження на акт метааргументації. Застосовуючи цю оптику, я класифікую мисленнєвий експеримент «Китайська кімната» як оціночну метааргументацію (evaluative meta-argument). Його головним висновком є оцінка можливості появи розуміння у сильного штучного інтелекту (Finocchiaro, 2013: p. 35).

Як наголошує Фінок'яро, критично важливою є ідентифікація конкретного опонента. Мішенню для Серля виступає не комп'ютер як інструмент, а амбітна претензія «сильного ШІ», згідно з якою програма не просто моделює розум, а буквально є ним. Ця метааргументація має цілком конкретну адресу: роботи Роджера Шенка. Принциповим є розрізнення історичного приводу та власне філософської мішені полеміки. Хоча текстуально критика 1980 року стосується скриптів Шенка, вони слугують для Серля лише «зручним прикладом» (Searle, 1980: p. 417), адже він підважує ширшу доктрину: обчислювальну свідомість (Computational Theory of Mind). Осягнути силу серлевого удару можна лише через реконструкцію цієї мішені.

Фундамент дискусії заклав ще Алан Тюринґ у своїй праці «Обчислювальні машини та інтелект». Він запропонував замінити туманне (на думку Тюринґа) питання про те, «чи можуть машини мислити» на прагматичну «гру в імітацію» (Turing, 1950: p. 433). Тюринґ робить заміну:

термін мислення (Thinking) через складну історію, метафізичний шлейф (в тому числі історичну пов'язаність із душею та внутрішніми процесами людини) та неможливість дати релевантне визначення замінене на термін інтелект (Intelligence). Інтелект в англійській мові має вужче значення, ніж мислення: це скоріше «кмітливість», здатність розв'язувати задачі, вміти адаптуватись. Тюринґ також не дає визначення цьому терміну, але відразу замінює його на властивість — розумність (Intelligent), яку в подальшому пропонує перевіряти вже відомою «грою в імітацію». Такий підхід був релевантним на той час через популярність біхевіоризму, який акцентував увагу на поведінці, а не на внутрішніх станах. Тому «інтелектуальна поведінка» (intelligent behaviour) суголосна з «грою в імітацію». На додачу до цього, маю припущення про вплив Віденського кола на Тюринґа та спробу використання їхнього підходу до розв'язання метафізичних (філософських) питань через аналіз мови. Остання теза потребує додаткового дослідження та (можливо) буде розкрита в наступних розвідках.

Теоретичним каркасом, що з високою ймовірністю дозволив перетворити гру Тюринґа на ідею «сильного ШІ», став обчислювальний функціоналізм. У 1960—70-х роках цей напрям набув статусу панівної парадигми у філософії свідомості, здійснивши радикальний методологічний розворот від фізичної реалізації мізків до функціональної організації процесу мислення. Це було важливим етапом у розвитку філософії свідомості, адже до цього панувала теорія тотожності роботи нейронів та свідомості (Mind). Головним дослідником тут виступив Гіларі Патнем, який у праці «Природа ментальних станів» (вперше опублікованій як «Психологічні предикати» у 1967 році) сформулював концепцію машинного функціоналізму: ментальні стани ідентичні не хімії мозку, а функціональним станам ймовірнісного автомата (Putnam, 1975: p. 434). У цій візії «розум» — це лише організаційна структура (функція), що співвідноситься з мозком так само, як програмне забезпечення з апаратним.

Ключовим наслідком теорії стає теза про «множинну реалізованість» (multiple realizability) ментального. Патнем наполягає: фізична природа носія функцій є абсолютно не важливою, адже ментальні властивості детермінуються суто функціональною роллю: каузальними зв'язками між вхідними стимулами, внутрішніми станами та поведінкою. У пізнішій статті «Філософія і наше ментальне життя» ця ідея стає всеосяжною та вважається розв'язком більшості проблем, в тому числі свідомості. Вдаючись до яскравої метафори Патнем стверджує, що людська істота могла б складатися з будь-чого (навіть зі «швейцарського сиру»), за умови належної організації матерії для підтримки функціональних станів (Putnam, 1975: p. 291). Така позиція фактично надала дослідникам ШІ карт-бланш: нівелювання значення матерії зрівнює в онтологічних правах кремнієвий чіп і біологічний нейрон. Єдиною умовою залишається коректне виконання алгоритму.

Систематизацію цієї ідеї здійснив Джері Фодор, розробивши обчислювальну теорію розуму (Computational Theory of Mind). Фундаментом мислення (за Фодором) є мова думки (mentalese) — внутрішня репрезентативна система з власним синтаксисом і семантикою. Когнітивні процеси тут постають як суто формальні операції над символами, де саме синтаксичні властивості репрезентацій визначають вміст ментальних процесів (Fodor, 1975: p. 34). Фодор формулює та висуває сміливе припущення: побудова машини, що маніпулює символами за правилами збереження істинності (truth-preserving rules), автоматично гарантує появу раціонального мислення. Фодор у своїй праці проводить паралелі та вказує на спорідненість інтелекту та раціонального мислення. Це підсилює мою здогадку про спорідненість та продовження ідей Тюринґа (Fodor, 1975: ch. 2). Цей непорушний альянс тез Патнема (про неважливість субстрату) та Фодора (про достатність синтаксису) дав можливість з'явитись концепції «сильного ШІ». Яку, своєю чергою, почав критикувати Серль.

Хвиля оптимізму щодо символьного підходу (Good Old-Fashioned AI) наприкінці 1970-х років спонукала дослідників до реалізації візії Тюринґа через створення програм для обробки природної мови. Саме в цей час публікуються роботи Роджера Шенка та його колег із Єльського університету, що стали безпосередньою мішенню серлевої критики. У монографії «Сценарії, плани, цілі та розуміння» автори розвивають теорію, де розуміння мови спирається на «сценарії» (scripts): попередньо визначені структури даних для опису стереотипних ситуацій (Schank & Abelson, 1977).

Програми на кшталт SAM (Script Applier Mechanism) мали на меті симуляцію розуміння історій через зіставлення тексту зі сценаріями. Хрестоматійний приклад, який аналізує Серль, — це сценарій відвідування ресторану. Машині пропонують історію про чоловіка, який замовив гамбургер, отримав підгорілу страву і пішов, не заплативши. На запитання: «Чи з'їв чоловік гамбургер?» — програма, керуючись алгоритмом сценарію, відповідає: «Ні». Прихильники «сильного ШІ» вбачали у здатності машини відповідати на питання про імпліцитну інформацію (відсутню в тексті) доказом наявності справжнього розуміння, еквівалентного людському (Searle, 1980: p. 417).

Мисленнєвий експеримент «Китайська кімната» спрямований саме проти цього епістемологічного стрибка: від успішного проходження спрощеного тесту Тюринґа до приписування машині ментальних станів. Серль ставив перед собою мету продемонструвати, що маніпуляція символами за правилами сценаріїв Шенка (попри зовнішню успішність) не породжує внутрішнього розуміння.

Сценарій мисленнєвого експерименту «Китайська кімната». Серль пропонує уявити гіпотетичну конструкцію, яка структурно відтворює роботу комп'ютера, але реалізується людиною в середині кімнати.

Задля забезпечення чистоти аргументації Серль моделює ситуацію, де процес обробки інформації стає доступним для безпосереднього людського спостереження. У центрі експерименту перебуває суб'єкт (сам автор), ізольований у замкненому просторі. Найважливішою передумовою повинна бути його лінгвістична обмеженість: будучи носієм англійської, він не володіє китайською мовою, сприймаючи ієрогліфи суто як набір позбавлених семантики (значень) графічних візерунків або «карлючок» (squiggles and squoggles).

Експериментальна процедура розгортається через послідовне надходження трьох пакетів вхідних даних: «сценарію», «історії» та, згодом, «питань» до тексту. Головним інструментарієм суб'єкта слугує книга правил англійською мовою з вичерпними інструкціями щодо маніпуляції символами. Важливий нюанс: правила оперують виключно формою символів, не посилаючись на їхнє значення. Типовий алгоритм виглядає гранично формально: «Побачивши у вхідному повідомленні символ форми 'X', візьміть символ форми 'Y' і покладіть його у вихідний слот» (Searle, 1980: p. 418).

Неухильне дотримання алгоритму дозволяє суб'єкту обробляти вхідні дані та відправляти повідомлення. Для зовнішніх спостерігачів (носіїв китайської мови) ці результати виглядають абсолютно релевантно, як осмислені відповіді. Точність інструкцій гарантує, що відповіді суб'єкта в середні кімнати залишаються зрозумілими та не відрізняються від відповідей реальної людини, яка вільно розмовляє мовою. Відповідно, система проходить тест Тюринґа: вона переконує спостерігача у наявності розумного співрозмовника. Водночас в середині кімнати ситуація протилежна: суб'єкт перебуває у стані повної когнітивної ізоляції, навіть не підозрюючи, що оперує мовою. Його механічні дії є відповідями на запитання про прочитану історію (Searle, 1980: p. 419).

Синтаксична досконалість та семантична порожнеча. Аналіз наслідків цього сценарію підводить Серля до спростування засновків сильного штучного інтелекту. Ключова проблема пролягає у відсутності можливого переходу між синтаксисом і семантикою. Суб'єкт у кімнаті досягає абсолютної синтаксичної досконалості: маніпуляції символами настільки точні, що дозволяють успішно пройти тест Тюринґа. Це не залишає зовнішнім спостерігачам жодного шансу вважати, що їм відповідав не носій мови. Однак, попри цю синтаксичну бездоганність, свідомість суб'єкта в середині кімнати фіксує цілковиту «семантичну порожнечу» стосовно китайської мови. Символи для нього залишаються суто зображеннями на картках, що ідентифікуються виключно за геометрією (shape). Це ще раз підтверджує, що вони для нього позбавлені будь-якого змісту, референції або інтенційності (Searle, 1980: p. 422).

Цей розрив слугує Серлю для демонстрації принципової тези: комп'ютерні програми за своєю сутністю є суто синтаксичними, оскіль-

ки визначаються через формальні операції над абстрактними символами. Людський розум, натомість, оперує не просто знаками, а ментальним змістом — семантикою. Розуміння передбачає знання того, що означає символ, на що він вказує у світі і яким є його предметне значення (денотат). Аргумент кристалізується у твердженні, що «синтаксис сам по собі не є достатнім для семантики» (syntax is not sufficient for semantics). Жоден обсяг формальних маніпуляцій (хоч би яким складним чи розгалуженим він був) не здатен самостійно породити значення (Searle, 1980: p. 419).

Такий висновок стає викликом для класичного функціоналізму: ідеальна кореляція між «входом» та «виходом» сама по собі нічого не гарантує. За правильною відповіддю може критися порожнеча: відсутність відповідного ментального стану. Причинно-наслідковий зв'язок тут такий: якщо людина-оператор, бездоганно виконуючи інструкції Шенка, не розуміє китайської мови, то немає жодних підстав приписувати це розуміння комп'ютеру, який так само сліпо слідує правилам. Серль констатує: комп'ютер володіє синтаксисом, але позбавлений семантики. Уявлення про «сильний ШІ» (Синтаксис + Апаратне забезпечення = Розум) відкидається як хибне. Оперування символами потребують інтерпретатора для набуття денотату. Самі по собі алгоритми неспроможні повноцінно породжувати семантику та розуміння, як це відбувається в біологічному мозку (Searle, 1980: p. 420).

Мисленнєвий експеримент Серля спровокував в академічному середовищі велику кількість публікацій в тому числі критичних робіт та заперечень висновків. Разом із оригінальною статтею в BBS опублікували коментарі 27 дослідників, на які були надані розлогі відповіді. Саме це обговорення дозволило прояснити основні напрямки захисту «сильного ШІ» та дозволило Серлю суттєво уточнити свою аргументацію.

У межах мого дослідження я аналізував лише ті контраргументи, які є суттєвими та становлять для аналізу експерименту найбільший теоретичний виклик. По-перше, на мою думку, ці заперечення репрезентують найсильніші позиції функціоналізму. По-друге, безпосередньо корелюють із критичними запитаннями (CQ) Дуґласа Волтона.

Рівень II. Прагма-діалектична динаміка

Одним із впливових підходів для аналізу теорії аргументації в сучасному контексті є прагма-діалектичний підхід. Використовуючи його для аналізу мисленнєвих експериментів, я буду реконструювати не так логічні конструкції, як динамічні комунікативні події. Найґрунтовнішу розробку цього підходу щодо «Китайської кімнати» здійснив Євген Попа у своєму дисертаційному дослідженні «Мисленнєві експерименти в академічній комунікації». Попередні дослідники схильні редукувати мисленнє-

вий експеримент Серля до статичної схеми засновків та висновків, наприклад, modus tollens (див.: Copeland, 2002; Jacquette, 1989; Hauser, 1997). Попа розглядає експеримент як складну мовленнєву комунікацію, що розгортається в конкретному інституційному полі та проходить чотири стадії критичної дискусії: конфронтацію, відкриття, аргументацію та завершення (Popa, 2016: p. 21).

Уже на самому початку цієї реконструкції «Китайської кімнати» (на стадії конфронтації) Попа фіксує момент визначення розбіжності в думках. Джон Серль тут виступає не відстороненим автором, а активним учасником — «рецензентом», що атакує позиції прихильників сильного ШІ. Паралельно з цим, у предметі суперечки виявляється прихована двозначність. На поверхні текстові маркери натякають на дискусію про комп'ютери (чи здатна машина розуміти), але Євген Попа під час аналізу доходить висновку про інший рівень суперечки. Дискусія відбувається навколо природи ментальних станів, а також питання: чи може уявлення про сильний ШІ вважатися коректним відображенням теорії свідомості (Popa, 2016: p. 41).

Попа акцентує увагу на стадії відкриття, де відбувається узгодження вихідних позицій. Знамените серлівське «уявіть собі, що я замкнений у кімнаті...» функціонує тут як мовленнєвий акт пропозиції (proposal). На самому початку мисленнєвого експерименту встановлюється рамка дискусії: перевірка тези щодо сильного ШІ відбуватиметься виключно крізь призму запропонованого сценарію. Опоненти часто потрапляють у цю пастку. Вступаючи в полеміку (наприклад, із «Системною відповіддю»), вони без заперечень приймають виклик, погоджуючись грати за правилами Серля. Альтернативною стратегією виступає хіба що «Відповідь робота». Попа інтерпретує її як спробу замінити підхід у цілому: намагання висунути контрсценарій і змінити вихідні параметри експерименту (Popa, 2016: p. 67).

На стадії аргументації сценарій використовується інструментально. На думку Попи, Серль не просто описує ситуацію, а виконує серію продуманих ходів. Вони покликані переконати авдиторію в неприйнятності наслідків теорії сильного ШІ. Специфіка підходу Попи полягає у врахуванні «стратегічного маневрування» учасників. З цим пов'язано балансування між діалектичною (раціональне вирішення суперечки) та риторичною цілями (перемога над опонентом). Частиною цього маневрування стає емоційно забарвлена лексика, як-от взаємні звинувачення в «метафізичних зобов'язаннях» (Popa, 2016: p. 12). Така перспектива дозволяє побачити «Китайську кімнату» як тривалий, незавершений діалог. У ньому кожен новий раунд критики стає черговим комунікативним ходом у спільній спробі вирішити (або поглибити) академічну розбіжність.

Встановлення історичного та дискурсивного контекстів дозволяє мені перейти до аналізу внутрішньої структури мисленнєвого експерименту. Чому порівняння «Китайської кімнати» з комп'ютером здається

мені коректним і де саме в ньому прихована пастка? Відповідь дає структурний аналіз аналогії на наступних двох рівнях. Я аналізую мисленнєвий експеримент як аргумент за аналогією, адже саме ця логічна форма найкраще демонструє механізм перенесення певного уявлення (властивість, функцію тощо) з гіпотетичного сценарію (джерела) на досліджувану проблему (ціль). Такий підхід дозволяє вийти за межі суб'єктивних інтуїцій і застосувати вже відомі методи. По-перше, перевірку структурного ізоморфізму (за Бартою) та, по-друге, діалектичної стійкості до контраналогій (за Волтоном).

Рівень III. Анатомія аналогії

Наступні два рівні будуть пов'язані з аналізом внутрішньої структури мисленнєвого експерименту та спробою оцінити її стабільність. Для першої частини цього аналізу я буду використовувати формальний апарат теорії аналогій Пола Барти. У своїй праці «Шляхом паралельних міркувань» (By Parallel Reasoning) Барта пропонує аналізувати аргументацію через розрізнення двох векторів: горизонтального та вертикального. Перші відображають подібність між властивостями джерела (source domain) та цілі (target domain). Інші описують внутрішню механіку: каузальні чи логічні ланцюжки всередині кожного домену окремо (Bartha, 2010: p. 13). Застосування цієї оптики до «Китайської кімнати» висвітлює принциповий нюанс: переконливість позиції Серля тримається на демонстрації розриву саме у вертикальних відношеннях. Хоча також можна констатувати очевидну горизонтальну симетрію.

Горизонтальний зріз експерименту справді демонструє майже ідеальний ізоморфізм між діями людини в кімнаті та роботою процесора. Маніпуляції суб'єкта з китайськими символами (переміщення карток, порівняння візерунків) функціонально тотожні операціям процесора, що виконує таку ж функціональну програму. Серль наголошує: людина тут перетворюється на живу реалізацію (instantiation) алгоритму, де вхідні дані, протоколи обробки та фінальний результат збігаються настільки бездоганно, що проходять тест Тюринґа (Searle, 1980: p. 418). Саме ця тотожність провокує перенесення властивостей за аналогією. Згода з тим, що комп'ютер крок у крок дублює дії людини, формує очікування ідентичних результатів, включно з розумінням.

Втім, зміщення фокусу на вертикальні відношення руйнує цю ілюзію, оголюючи критичну дисаналогію в каузальній структурі. Біологічний мозок, що виступає прихованим еталоном, характеризується прямим вертикальним зв'язком: нейрофізіологія тут безпосередньо породжує ментальний зміст (семантику). За Серлем, мозок володіє специфічними «каузальними силами» (causal powers), що й забезпечують «розуміння» (Searle,

Таблиця 1. **Вертикальні та горизонтальні подібності між «Китайською кімнатою» та комп'ютером**

Вертикальні Подібності	Домен Кімната з Серлем (С)	Горизонтальні Подібності	Домен Комп'ютер (Т)
↓↑	Наповнена коробками з китайськими символами	← →	Має базу даних
	Має книжку з інструкціями щодо роботи з символами	← →	Має програму
	Отримує запитання китайською мовою	← →	Має пристрій введення
	Надає правильні відповіді на запитання китайською мовою	← →	Має пристрій виведення
	Подібність, що допускається:		
	Відсутнє розуміння (семантичний рівень)	⇒	Можливо, відсутнє розуміння (семантичний рівень)

1980: p. 422). Натомість модель «Китайської кімнати» демонструє відсутність вертикального зв'язку між формальною маніпуляцією символами (синтаксисом) та розумінням змісту (семантикою). Бездоганне виконання синтаксичних правил не наближає людину до розуміння мови. Отже, синтаксис залишається герметично ізольованим від семантики.

Методологія Барти дозволяє кваліфікувати цю ситуацію як «розрив передбачення». Оскільки вертикальне відношення «синтаксис/семантика» не працює в джерельному домені (кімната), зникають логічні підстави прогнозувати його наявність у цільовому домені (сильний ШІ) (Bartha, 2010: p. 253). Аналіз експерименту доводить: горизонтальної подібності алгоритмів критично недостатньо для обґрунтування тотожності когнітивних станів. ШІ здатен відтворити лише синтаксичний вимір процесу, паралельно ігноруючи його каузальний біологічний фундамент та не мати семантичного виміру (за класичним поділом семіотики) (див. табл. 1).

Можливість фіксації «розриву» у вертикальних відношеннях безпосередньо залежить від прийняття однієї умови: серлевої передумови біологічного натуралізму. Цей момент функціонує як засновок: мозок володіє специфічними «каузальними силами», яких кремній позбавлений. Відмова від цієї метафізичної передумови (шлях функціоналістів) автоматично робить вертикальну структуру програми цілком ізоморфною до структури розуму. У цьому й полягає евристична цінність методу Барти: не створюючи нових сутностей, він експлікує цей засновок.

Структурний аналіз сам по собі не доводить наявності у мозку унікальних «каузальних сил». Він демонструє інше — структурну залежність аргументації від прийняття сильного метафізичного припущення. Варто відмовитися від такого припущення, що «біохімія має значення»,

як вертикальний розрив в аналогії зникає, позбавляючи аргумент передбачувальної сили. Використання методу Барти виявляє, що «Китайська кімната» є не так спробую спростування сильного ШІ, як демонстрацією фундаментальної несумісності функціоналізму з інтуїціями про біологічну реалізацію свідомості.

Хоча аналіз дозволяє продемонструвати можливість «вертикального розриву» між синтаксисом і семантикою, в академічному контексті будь-яка аналогія залишається лише правдоподібною, а відтак — спростовною (defeasible). Потрібно оцінити стійкість конструкції аналогії, що потребує від мене спроби перевірки на діалектичну міцність (robustness). Саме для цього я застосую систему критичних запитань Дуґласа Волтона.

Рівень IV. Діалектична стійкість: Схеми та критичні запитання

Застосування теорії аргументаційних схем Дуґласа Волтона змінює саму оптику дослідження. Аналіз «Китайської кімнати» переміщується з площини формальної логіки в простір діалектики. Якщо підхід Пола Барти перевіряє структурну міцність аналогії, то Волтон пропонує розглядати мисленнєвий експеримент передусім як хід у діалозі. Його тактична мета — встановити презумпцію і, що принципово, перекласти тягар доведення на опонента.

У класифікації Волтона, Рида та Маканьо конструкція Серля ідентифікується як «аргумент від аналогії» (Argument from Analogy). Це важливий нюанс. Належність до класу спростовних міркувань (defeasible reasoning) означає, що висновок не випливає через логічне слідування з засновків. Він залишається правдоподібним лише доти, доки немає доказів протилежного. Сила аргументації залежить не від логічного слідування, а від того, чи вистоїть пропонент під градом специфічних «критичних запитань» (Walton, Reed, & Macagno, 2008: p. 315).

Механіка схеми розгортається через зіставлення. Є джерельний випадок (C_1) — та сама гіпотетична людина, яка перекладає картки в кімнаті. І є цільовий випадок (C_2) — цифровий комп'ютер, запрограмований за канонами сильного ШІ. Аргументація тримається на двох опорних точках. Перша — «засновок подібності» (Similarity Premise): він фіксує суттєву близькість C_1 і C_2 в аспекті виконання формальних операцій. Друга — «базовий засновок» (Base Premise): ми приймаємо як факт, що в кімнаті (C_1) істинним є твердження A (розуміння китайської мови відсутнє). На цій підставі формується висновок, що твердження A істинне і для комп'ютера (C_2) (Walton et al., 2008: p. 58). Цей маневр не ставить крапку в пошуках істини, проте створює вагому презумптивну перевагу на користь Серля.

Така конфігурація змушує опонентів (прихильників сильного ШІ) змінювати стратегію. Просто відкинути висновок уже недостатньо: доводиться йти в контратаку на засновки через систему критичних запитань (Critical Questions).

Аналіз літератури виявляє показову деталь: ключові історичні заперечення проти «Китайської кімнати» фактично є розгорнутими відповідями на запитання Волтона. Скажімо, питання про те, «чи існують між C_1 і C_2 відмінності, які підривали б силу аналогії», прямо корелює з «Відповіддю симулятора мозку» та «Відповіддю робота». Обидві вказують на те, чого бракує кімнаті, але не мозку чи роботу — каузального зв'язку із зовнішнім світом. Інше питання б'є ще точніше: «Чи є схожість у маніпуляції символами достатньою підставою для перенесення властивості "нерозуміння"?» Це, по суті, квінтесенція «Системної відповіді», яка заперечує релевантність досвіду окремого процесора для оцінки свідомості системи в цілому (Searle, 1980: p. 419). Метод Волтона дозволяє впорядкувати цей хаос, перетворюючи полеміку на структурований стрес-тест для аналогії (див. табл. 2).

Системна відповідь (The Systems Reply). Мабуть, найсерйознішим інтелектуальним викликом для аргументу Серля залишається «Системна відповідь». Цей контрприклад виник ще під час ранніх дискусій у Берклі і радикально змінив аналітичну оптику. Фокус тут зміщується з індивідуального суб'єкта на систему в цілому.

Адепти цієї позиції охоче погоджуються: так, людина, замкнена в кімнаті, китайської не розуміє. Але для них цей факт є тривіальним. Він нічого не спростовує. Людина тут виконує суто технічну функцію — це, по суті, лише центральний процесор (CPU).

Фундамент аргументації — голізм. Розуміння приписується не ізольованому елементу, а сукупності взаємодіючих частин. Система інтегрує не лише оператора, а й книгу правил (програму), набори символів (дані), кошики для сортування і навіть фізичний простір кімнати. Критики наполягають: відсутність семантики у індивіда не означає, що «система як ціле не розуміє» (the system as a whole understands) (Searle, 1980: p. 419). Найкращою ілюстрацією тут слугує аналогія з мозком. Жоден окремий нейрон не володіє свідомістю, проте мозок як цілісність ці властивості демонструє. За такою логікою, когнітивна сліпота оператора не може бути доказом того, що система, яка успішно проходить тест Тюринґа, теж «сліпа».

Контраргумент Серля: Інтерналізація системи. Щоб спростувати цей аргумент, Серль модифікує початковий експеримент. Його відповідь спирається на концепцію «інтерналізації» (internalization). Автор пропонує сценарій, де суб'єкт не просто гортає книгу правил, а вивчає її напам'ять. Ба більше, він може вийти з кімнати на відкритий простір і виконувати всі операції подумки. У такій конфігурації людина

Таблиця 2. **Діалектична реконструкція «Китайської кімнати» за схемою Д. Волтона**

Елемент схеми аргументації	Формальний опис (Walton, Reed, & Macagno, 2008)	Реалізація в аргументі Дж. Серля (Searle, 1980)
Засновок подібності (Similarity Premise)	Як правило, випадок C_1 подібний до випадку C_2	Дії людини в кімнаті (C_1), яка виконує інструкції з маніпуляції символами, функціонально подібні до роботи комп'ютера (C_2), що виконує програму Шенка. В обох випадках відбувається суто синтаксична обробка даних
Базовий засновок (Base Premise)	Твердження A є істинним (або хибним) у випадку C_1	Людина в кімнаті (C_1) не розуміє китайську мову (A є істинним). Вона не знає значення символів, якими маніпулює
Висновок (Conclusion)	Твердження A є істинним (або хибним) у випадку C_2	Отже, комп'ютер (C_2), виконуючи програму, також не розуміє китайську мову (A є істинним для C_2)
Критичне запитання 1 (CQ1: Differences)	Чи існують відмінності між C_1 та C_2, які б підривали силу подібності?	Атака опонентів («Відповідь симулятора мозку»): Так. Мозок (C_2 в ідеальній симуляції) має складну нейронну структуру та каузальні зв'язки, яких немає у людини з картками (C_1)
Критичне запитання 2 (CQ2: Relevance)	Чи є подібність у зазначеному аспекті релевантною для перенесення властивості A?	Атака опонентів («Системна відповідь»): Ні. Той факт, що частина системи (людина) не розуміє, не є релевантним для висновку, що вся система (C_2) не розуміє
Критичне запитання 3 (CQ3: Counter-analogy)	Чи існує інший випадок C_3, який також подібний до C_1, але в якому A є хибним?	Можлива атака: Чи можемо ми уявити систему, яка маніпулює символами, але яку ми все ж вважаємо розумною? (Наприклад, нейронна мережа людського мозку, яка теж «сліпо» передає сигнали)

буквально вбирає в себе всю систему. Вона стає водночас процесором, програмою та сховищем даних.

Міркування Серля будується на тотожності: якщо людина інтерналізувала систему, відмінність між ними зникає. Утім, навіть ставши системою, суб'єкт не починає розуміти китайську. Він продовжує маніпулювати за тим само алгоритмом. Серль резюмує жорстко: «оскільки людина не розуміє, то й система не розуміє, бо в системі немає нічого, чого б не було в людині» (Searle, 1980: p. 420). Цим він намагається довести, що навіть повна інтерналізація синтаксису не породжує семантику. Проблема не в масштабі (частина чи ціле), а в природі процесу: формальні маніпуляції, де б вони не відбувалися, не здатні створити семантику.

Аналіз через схеми Волтона. Оскільки «Системна відповідь» не змогла переконати всіх остаточно, критики — зокрема Нед Блок (Block, 1980)

та Зенон Пилішин (Pylyshyn, 1980) — пішли далі, сформулювавши аргумент «Симулятора мозку».

Формалізувати цей діалектичний поворот дозволяють схеми Волтона. По суті, «Системна відповідь» атакує аргумент через критичне запитання щодо композиції (Composition CQ): «Чи є відсутність властивості (розуміння) у частини підставою заперечувати цю властивість у цілого?». Стратегія інтерналізації слугує Серлю блокуванням цього питання. Він просто ототожнює частину і ціле.

Проте уважний аналіз виявляє тут вразливість. Захист Серля працює лише за умови повного ігнорування ефекту емерджентності. Навіть якщо «людина-система» інтроспективно не фіксує розуміння, це залишає відкритим інше критичне запитання (Counter-Analogy CQ). Чи може розуміння виникати як емерджентна властивість, що принципово недоступна свідомості процесора — так само, як нейрон не «відчуває» думок мозку? Метод Волтона показує: «перемога» Серля над «Системною відповіддю» тримається на феноменологічному критерії (відсутність відчуття). А це критерій, який функціоналісти приймати зовсім не зобов'язані.

Відповідь симулятора мозку (The Brain Simulator Reply). Ця лінія аргументації підвищує ставки. У термінах схеми Волтона вона актуалізує критичне запитання про відмінності (CQ1: Differences): чи можна будувати правильну аналогію на основі маніпуляції символами, якщо ми при цьому повністю ігноруємо фізіологію мозку?

Аргументація тут така. Оскільки попередні заперечення не змогли похитнути позицію Серля щодо «сценарної» обробки символів, критики (насамперед дослідники з Берклі та M.I.T) вдалися до фундаментальнішого підходу. Це і є «Відповідь симулятора мозку». Суть стратегії полягає у різкій зміні рівня абстракції. Ми припиняємо розглядати програму як набір інструкцій для розуміння тексту (як у кейсі зі сценаріями Шенка) і спускаємося до рівня «заліза» — до моделювання фізичної архітектури мозку.

Гіпотеза смілива. Вона передбачає створення програми, що симулює не абстрактний інформаційний потік, а реальну послідовність нейронних збуджень у синапсах. Програма має точно відтворити момент розуміння і формування відповіді на нейрофізіологічному рівні.

У такій моделі комп'ютер виконує потрійну роботу: інтегрує вхідні дані (історії та питання), симулює формальну структуру нейронної активності «Китайського кімнати» й видає відповідні символи на виході. Позиція прихильників цього підходу категорична. Якщо машина функціонально дублює діяльність мозку (який, безперечно, розуміє), то заперечувати наявність розуміння у машини — це впадати в логічну суперечність. Адже на рівні синапсів функціональна різниця зникає (Searle, 1980: p. 419). «Відповідь симулятора мозку» прямо апелює до головної тези функціоналізму:

ментальні стани — це лише каузальні ролі. Відтворіть роль, навіть через симуляцію: ви гарантовано отримаєте ментальний стан.

Відповідь робота (The Robot Reply). Якщо попередня атака йшла через «мозок», то ця заходить з боку «тіла». Критика тут зміщує акцент на аспект каузального зв'язку (CQ1/CQ2) і фокусується на головній ваді «Кімнати» — відсутності сенсорів та ефекторів. Без них символи не мають шансів «заземлитися» в реальності.

Нездатність «Системної відповіді» розсіяти сумніви спонукала єльських критиків (зокрема Роджера Шенка (Schank, 1980) та Роберта Віленскі (Wilensky, 1980)) сформулювати альтернативу. Так з'явилася «Відповідь робота». Ця позиція фактично визнає поразку ізольованого комп'ютера. Справді, замкнене в чотирьох стінах оперування символами залишається самореферентним; це гра слів без виходу до реальності. Стратегія порятунку —інтеграція. Комп'ютер (або людину в кімнаті) слід помістити в тіло фізичного робота. Така система отримує не лише процесор, а й «очі» (камери) для сприйняття довкілля та «руки» (двигуни, маніпулятори) для фізичної інтервенції у світ (Searle, 1980: p. 420).

Фундаментом тут слугує концепція каузального заземлення (causal grounding). Прихильники концепції сильного ШІ наполягають: все змінюється, коли робот здатен сприйняти візуальний образ (наприклад, гамбургер), розпізнати його і фізично з'їсти. Ці дії трансформують онтологічний статус символів у його «мозку». Вони перестають бути абстракцією. Набуття реального семантичного змісту забезпечується тут прямими каузальними ланцюгами з фізичними об'єктами. Розуміння в такій архітектурі — це не наслідок перестановки знаків, а результат динамічної інтеракції програми із зовнішнім середовищем. Саме це, на думку критиків, гарантує омріяний перехід від синтаксису до семантики (розуміння).

Контраргумент Серля: Розширення формалізму. Як відповідає на цей виклик Серль? Він доводить: додавання «перцепції» та «дії» не змінює суті справи. Обробка інформації залишається просто обробкою інформації. Щоб спростувати «Відповідь робота», він знову модифікує експеримент, поміщаючи суб'єкт (себе) безпосередньо в черепну коробку робота.

У цьому сценарії людина перебирає на себе функцію центрального процесора і керує всім «тілом». Вхідні дані надходять уже не через двері у вигляді карток, а як потік символів від сенсорів — камер і мікрофонів. Вихідні ж знаки, які генерує суб'єкт, активують двигуни. Критично важливий нюанс: оператор усередині й гадки не має про природу цих знаків. Чи це дані з камери? Чи команда для моторів? Для нього це, як і раніше, лише невідомі «карлючки», маніпуляція якими суворо регламентована книгою правил (Searle, 1980: p. 420).

Серль аргументує: навіть прямий контакт зі світом через тіло робота не дарує суб'єкту розуміння. Уявімо, що робот «бачить» гамбургер. Лю-

дина всередині при цьому споглядає лише черговий символ — скажімо, хвилясту лінію. Виконуючи інструкції, вона видає інший знак. Цей сигнал змушує руку робота рухатися і хапати їжу. Проте суб'єкт не усвідомлює ні гамбургер, ні руку, ні сам факт руху. Весь цей інформаційний потік залишається для нього процесом суто алгоритмічним (синтаксичним).

Так Серль демонструє фундаментальну межу. Ви можете нарощувати зовнішні зв'язки («перцептивні» чи «моторні»), але це лише збільшує обсяг вхідних символів. Вони не трансформуються у ментальний зміст (розуміння). Більше синтаксису не народжує семантику. Робот, подібно до кімнати, залишається системою, де «всередині темно»: там панує сліпе виконання алгоритмів, позбавлене будь-якого усвідомлення та референції до реального світу.

Метод критичних запитань чітко фіксує: Серль успішно заблокував атаки на окремі компоненти експерименту (CQ1 та CQ2). Усвідомлення цього факту змушує критиків змінити тактику. Їм залишається вдатися до стратегії кумуляції — спроби посилити аргументацію, поєднавши всі фактори в єдину систему.

Комбінована відповідь (The Combination Reply). Оскільки жодне з попередніх заперечень (чи то системне, чи робототехнічне, чи симуляційне) не змогло самотужки порушити захист Серля, критики змінили тактику. Представники шкіл Стенфорда та Берклі вдалися до «стратегії кумулятивного ефекту». Ідея проста: якщо аргументи не працюють поодинці, варто об'єднати їх у єдину суперсистему. Сценарій має такий вигляд. Уявіть робота, якого візуально не відрізнити від людини. Він має повний арсенал сенсорів та ефекторів (відповідь робота). Але керує ним не простий скрипт, а комп'ютер, що виконує програму повної симуляції нейронної активності людського мозку (відповідь симулятора). Опоненти переконані: ця система поводилася б так само, як людина (системна відповідь), і відповідала б усім критеріям розумної істоти.

Ставку тут роблять на емерджентність. Нехай окремі компоненти («залізо», синапси, алгоритми) не гарантують розуміння, але їхня синергія, мовляв, породжує нову якість. Критики тиснуть на інтуїцію: зустріч із таким андроїдом змусила б нас приписати йому ментальні стани та відповідно розуміння. Інакше довелося б сумніватися і в людях, адже наша впевненість у чужому розумі спирається саме на поєднання адекватної поведінки та біологічної структури. Відмова визнати цю систему розумною виглядала б як прояв невиправданого «біологічного шовінізму» (Searle, 1980: p. 421).

Відповідь Серля: Нуль плюс нуль. Як відповідає на цей комплексний виклик Серль? Він доводить, що просте сумування синтаксичних процесів (незалежно від масштабу та складності) не здатне породити семантичний зміст. Для цього він модифікує експеримент востаннє, переміщуючи «Китайську кімнату» безпосередньо в череп робота.

У цій версії сам Серль виконує функцію центрального процесора. Сидячи всередині голови робота, він керує машиною, виконуючи інструкції тієї само програми-симулятора нейронів. Вхідні дані тепер надходять не через двері у формі карток, а прямо від телекамер та мікрофонів. Але для Серля це лише потік формальних символів. Він обробляє їх згідно з правилами і видає інші символи, що запускають мотори та голосовий апарат.

Результат для сильного ШІ виявляється нищівним (на думку Серля). Робот може поводитися як завгодно розумно, вести філософські диспути чи писати сонети. Але гомункул у його голові (Серль), який виконує всю когнітивну роботу, не розуміє суті процесу. Він лише перекладає знаки. Серль не знає, що вхідний символ означає «бачу червону квітку», а вихідний є командою «простягнути руку». Увесь процес залишається абстрактною грою. Цим Серль демонструє, що навіть тотальна мобілізація всіх стратегій ШІ не вирішує проблеми інтенціональності. Якщо жодна з частин системи (симуляція, робототехніка, алгоритми) не має доступу до семантики, то їхнє поєднання також залишається сліпим. Нуль, доданий до нуля, дає нуль. Чистий синтаксис, хоч у які складні фізичні оболонки його загортай, ніколи не трансформується у семантичне розуміння (Searle, 1980: p. 422).

Відповідь інших умів (The Other Minds Reply). На цьому етапі дискусія виходить на третє критичне запитання (CQ3: Counter-Analogy). Критики намагаються знайти контрприклад — ситуацію, у якій ми приписуємо розуміння виключно на основі поведінки, не маючи жодного доступу до внутрішніх станів. І такий приклад існує. Це інші люди.

Вагомим викликом (зокрема з боку єльської групи) стає «Відповідь інших умів». Вона експлуатує класичну епістемологічну проблему: ми не маємо прямого доступу до суб'єктивного досвіду іншої істоти. Прихильники цієї позиції ставлять питання: на якій підставі ми взагалі вирішили, що люди довкола нас — свідомі? Єдиним критерієм тут слугує поведінка. Якщо суб'єкт діє адекватно і реагує на стимули (як ми), то ми визнаємо його розумним.

Оскільки у випадку з комп'ютером нам теж доступна лише поведінка (вхід/вихід), застосування якихось інших, «завищених» стандартів скидалося б на дискримінацію. Ця модифікація мисленнєвого експерименту змушує нас визнати: якщо машина проходить поведінковий тест на рівні людини, ми зобов'язані приписати їй розуміння. Інакше, щоб бути послідовними, доведеться сумніватися в розумності всіх своїх знайомих (Searle, 1980: p. 421).

Контраргумент: Епістемологія проти онтології. Серль парирує цей закид, чітко розмежовуючи домени. Проблема інших умів є епістемологічною (як ми «дізнаємося» про стани), тоді як дискусія про силь-

ний ШІ — онтологічною (чим ці стани «є»). У реальному житті наша впевненість у чужій свідомості спирається не лише на поведінку, а й на презумпцію спільної біології. Ми припускаємо, що схожа «апаратура» (мозок) продукує схожі ефекти. Втім, аргумент «Інших умів» «б'є повз ціль» навіть без прив'язки до біології. У «Китайській кімнаті» питання стоїть не про те, як система виглядає для зовнішнього спостерігача (його вердикт про бездоганну поведінку ніхто не оскаржує). Питання в сутності процесів усередині.

У випадку з комп'ютером двозначність зникає. Ми достеменно знаємо, що відбувається всередині: маніпуляція формальними символами. Знаємо, бо самі це запрограмували. На відміну від ситуації з людьми, тут ми володіємо «інсайдерською інформацією». Ми знаємо про відсутність ментальних станів, бо суб'єкт-реалізатор системи (людина в кімнаті) їх не відчуває. Серль іронічно підсумовує позицію критиків: оскільки існує ризик помилки щодо свідомості людей (через брак знань), вони пропонують свідомо помилитися і щодо комп'ютерів — приписати їм розуміння всупереч «наявному» знанню про його відсутність. Поведінкова подібність втрачає доказову силу, коли структурний аналіз викриває порожнечу інтенціональності (Searle, 1980: p. 422).

Підсумки аналізу IV рівня. Застосування схем Волтона дозволяє фіналізувати статус аргументу. Серль доволі успішно блокує атаки на композицію («Системна відповідь») та контраналогії («Інші уми»). Його головна зброя — введення додаткових умів: інтерналізації (я стаю системою) та привілейованого доступу від першої особи (я знаю, що не розумію).

Проте аналіз виявляє серйозну аргументативну вразливість захисту щодо першого питання (CQ1: Відмінності). Перемога Серля тримається на прийнятті засновку біологічного натуралізму («мозок має специфічні каузальні сили, яких немає у силікону»). Якщо опонент відкидає цю аксіому, аргумент втрачає фундамент, адже зникає єдиний критерій принципової відмінності «кімнати» від «мозку».

Отже, «Китайська кімната» доводить не абсолютну неможливість штучного інтелекту як такого. Вона доводить його неможливість лише у межах суто синтаксичної онтології.

Висновки: Онтологічний статус програм та межі функціоналізму

Аналіз критичних заперечень крізь призму схем аргументації виявляє у полеміці навколо «Китайської кімнати» парадоксальну річ. Намагаючись розбити аргументацію Серля, опоненти часто (хоч і неявно) порушують власні принципи. Фактично вони відмовляються від фундаменту своєї ж позиції: доктрини комп'ютаційного функціоналізму.

Згадаймо вихідну точку. Базою для «сильного ШІ», яку атакував Серль, слугувала теза Гіларі Патнема про «множинну реалізованість» (multiple realizability) (Putnam, 1975). Цей підхід встановлює єдиний, жорсткий критерій наявності розуму: успішне виконання функції. Важливий лише алгоритм переходу від вхідних даних до вихідних. Фізична природа носія? Швидкість обробки? Складність каузальних зв'язків? Для чистого функціоналізму все це несуттєві деталі. Серль у своєму експерименті послідовно дотримується цієї умови. Система в кімнаті функціонально повністю еквівалентна комп'ютеру: за умови ідентичних вхідних даних вона генерує ідентичні вихідні.

Проте, зіткнувшись з інтуїтивною очевидністю того, що оператор кімнати нічого не розуміє, критики починають шукати рятівне коло в аргументах, що суперечать духу функціоналізму. Вони вказують на «надто повільну швидкість» роботи людини з картками. Або наголошують на відмінності у «матеріалі» (біологічний мозок проти кремнію). Цим опоненти (свідомо чи ні) вводять нові критерії: темпоральні, фізичні або структурні. Але ж початкове визначення сильного ШІ їх прямо не включало. Як слушно зауважує Серль, якщо функціоналізм істинний, параметри швидкості чи хімічного складу не повинні мати жодного впливу на онтологний статус ментального, відповідно розуміння (Searle, 1980: p. 422).

Відтак значна частина критики «Китайської кімнати» б'є повз ціль. Вона атакує аналогію на підставі ознак, які сам «сильний ШІ» оголосив іррелевантними. Це діалектична пастка. Коли критик стверджує, що кімната не розуміє, бо вона «не справжній мозок» або «недостатньо складна система», він фактично погоджується з Серлем. Він визнає: самого лише виконання формальної функції (синтаксису) для виникнення розуміння недостатньо. Цей поворот показовий. Експеримент досяг мети не так через пряме доведення неможливості сильного ШІ, як через викриття внутрішньої непослідовності функціоналістської парадигми. Щоб врятуватися, її послідовники змушені апелювати до позафункційних властивостей.

По-перше, Серль зробив фундаментальну річ: він чітко зафіксував розрив між синтаксисом і семантикою. Навіть якщо припустити, що цю прірву колись вдасться подолати (на що сподіваються прихильники системного підходу), правила гри змінилися. Тягар доведення відтепер лежить на опонентах. Саме їм належить пояснити механіку перетворення сухих формальних маніпуляцій на розуміння, а не просто постулювати це як аксіому.

По-друге, філософ повернув у центр дискусії біологію. Його концепція «біологічного натуралізму» трактує свідомість як каузально емерджентну властивість мозку: точно так само, як фотосинтез є властивістю

рослин. Це диктує чітку умову. Хочете створити штучну свідомість? Тоді доведеться дублювати каузальні сили мозку, а не просто копіювати його інформаційну архітектуру.

Сьогодні, в епоху великих мовних моделей (GPT, Claude чи Gemini тощо), аргумент Серля продовжує існувати як релевантний. Він має несподіване втілення у концепції «стохастичних папуг» (stochastic parrots). Цей термін, який ввели Емілі Бендер та Тимніт Гебру, влучно схоплює сутність сучасних систем. Це сутності, які «навмання зшивають разом послідовності мовних форм.., згідно з імовірнісною інформацією.., але без будь-якого посилання на значення» (Bender et al., 2021). Власне, феномен LLM стає ідеальним полігоном для перевірки мого методу, але висновок тут не буде чорно-білим.

У площині горизонтальних відношень (за Бартою) моделі сягнули стелі: імітація діалогу тут майже досконала. Класичний тест Тюринґа у його первісній біхевіористичній версії більше не працює як діагностичний інструмент. Проте текстові моделі, відрізані від сенсомоторного контакту зі світом, демонструють важливу річ: синтаксис здатен імітувати семантику до ступеня нерозрізненості. Це парадоксальним чином підтверджує тезу Серля. Успішна поведінка (passing the test) не гарантує «заземлення» (grounding), хоча й не виключає наявності складної внутрішньої структури.

Структурний аналіз, однак, підтримує висновки Серля. LLM залишаються в'язнями «кімнати тексту». Вони віртуозно оперують імовірнісними зв'язками між символами (інференційна семантика, але у термінології Серля це все одно синтаксис), але не мають доступу до референції — прямого каузального зв'язку з об'єктами світу (що у термінології Серля відповідає семантиці).

Тож ефект «стохастичних папуг» не просто ілюструє інтуїцію Серля, а уточнює межі її застосування. «Китайська кімната» успішно блокує претензії ШІ на людську інтенційність (grounded understanding). Втім, вона залишає простір для концепції функціонального розуміння, яке, можливо, не вимагає біологічної основи.

Застосування мого чотирирівневого аналізу дозволило вийти за межі класичної суперечки «інтуїція проти обчислень». Поєднання мета-аргументації (М. Фінок'яро) з прагма-діалектикою (Є. Попа) дозволило уточнити, що це динамічний інструмент, який ефективно перекладає тягар доведення на техноoптимістів. Відтак перевірка через критичні запитання Волтона та теорію аналогій Барти довела, що відповідь на питання про природу ШІ не може бути суто емпіричною. Все залежить від вибору базової аналогії. Чим є нейромережа: подобою мозку чи все ж таки книгою правил?

ДЖЕРЕЛА / REFERENCES

Bartha, P.F.A. (2010). *By parallel reasoning: The construction and evaluation of analogical arguments*. Oxford University Press.

Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610—623). ACM. https://doi.org/10.1145/3442188.3445922

Churchland, P.M. & Churchland, P.S. (1990). Could a machine think? *Scientific American*, 262(1), 32—39.

Copeland, B.J. (2002). The Chinese Room from a logical point of view. In: J. Preston & M. Bishop (Eds.), *Views into the Chinese room: New essays on Searle and Artificial Intelligence* (pp. 109—123). Clarendon Press.

Finocchiaro, M.A. (2005). *Arguments about Arguments: Systematic, Critical, and Historical Essays in Logical Theory*. Cambridge University Press.

Finocchiaro, M.A. (2013). *Meta-argumentation: An Approach to Logic and Argumentation Theory*. College Publications.

Fodor, J.A. (1975). *The language of thought*. Harvard University Press.

Hauser, L. (1997). Searle's Chinese Box: Debunking the Chinese Room Argument. *Minds and Machines*, 7(2), 199—226

Jacquette, D. (1989). Adventures in the Chinese Room. *Philosophy and Phenomenological Research*, 49(4), 605—623.

Olmos, P. (Ed.). (2017). *Narration as Argument*. Springer International Publishing.

Popa, E.O. (2016). *Thought experiments in academic communication: A pragma-dialectical method for reconstructing the argumentative use of imaginary scenarios in academic disputes.* [Doctoral dissertation, University of Amsterdam]. UvA-DARE (Digital Academic Repository).

Putnam, H. (1967). Psychological predicates. In: W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37—48). University of Pittsburgh Press.

Putnam, H. (1975). Philosophy and our mental life. In: *Mind, language and reality: Philosophical papers* (Vol. 2, pp. 291—303). Cambridge University Press.

Putnam, H. (1975). The Nature of Mental States. In: *Mind, Language and Reality: Philosophical Papers* (Vol. 2, pp. 429—440). Cambridge University Press.

Schank, R.C. & Abelson, R.P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates.

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417—424.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433—460. https://doi.org/10.1093/mind/LIX.236.433

Walton, D.N. & Krabbe, E.C.W. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Отримано / Received 24.10.2025
Прийнято до друку після рецензування /
Accepted for publication after review 09.01.2026
Підписано до друку / Signed for printing 02.04.2026

Ruslan MYRONENKO,
Master of Philosophy,
Founder of the educational project "Plato's Cave,"
6/11 Yevheniia Miroshnychenko St.,
Kyiv 03057, Ukraine;
PhD student (postgraduate researcher),
Department of Logic, Faculty of Philosophy,
Taras Shevchenko National University of Kyiv,
64/13 Volodymyrska St., Kyiv 01601, Ukraine;
Lead Engineer, Department of Social Philosophy, H. S. Skovoroda Institute of Philosophy of the National Academy of Sciences of Ukraine
mironrus@gmail.com
https://orcid.org/0000-0003-4058-9772
SCOPUS ID: 57206845047

THE ARCHITECTONICS OF THE «CHINESE ROOM»:
RECONSTRUCTION AND EVALUATION OF THE THOUGHT EXPERIMENT VIA ARGUMENTATION THEORY

The article analyzes John Searle's famous «Chinese Room» thought experiment, which denies the ability of artificial intelligence to achieve genuine understanding. The study aims to deconstruct this argumentation step by step to clarify exactly how it is structured, what makes it persuasive, and where its weaknesses lie. For this purpose, a model is applied that examines the text at four levels. The first level clarifies the historical background: specifically, whose ideas and computer programs the author opposed. The second level demonstrates how Searle constructs the experiment's scenario to subtly impose discussion rules that favor his position on the audience. In the third stage, the relevance of comparing a computer's operation to the actions of a person mechanically sorting incomprehensible characters is examined. The analysis reveals that the persuasiveness of the argumentation relies heavily on the assumption that only a biological brain is capable of generating consciousness. At the fourth level, the experiment is tested for robustness using the most well-known objections from critics. The conducted analysis shows that Searle successfully proves a machine's inability to understand meaning solely through the mechanical manipulation of symbols. However, his thought experiment proves vulnerable to the assumption that consciousness can emerge as an entirely new property of highly complex systems. Using modern Large Language Models (LLMs) as an example, the study concludes that the «Chinese Room» argument remains relevant, as it proves that machines do not grasp the physical connection between words and the real world. At the same time, this does not rule out the possibility that artificial intelligence is capable of successfully operating with syntactic connections between words within the language itself.

Keywords: *Chinese Room, John Searle, argumentation theory, philosophy of mind, informal logic, rhetoric, thought experiment, analogical reasoning, computational functionalism, meta-argumentation, pragma-dialectics, Strong AI, inferential semantics, LLM.*