

ПРОБЛЕМА АТРИБУЦІЇ ВІДПОВІДАЛЬНОСТІ: ІНСТРУМЕНТ, АГЕНТ, СУБ'ЄКТ

Малєєв А. К.,

аспірант

Київський національний університет імені Тараса Шевченка

Київ, Україна

У роботі досліджується проблема атрибуції відповідальності штучному інтелекту через розмежування трьох статусів штучної системи — інструмента, агента та суб'єкта відповідальності. Запропоновано розрізняти атрибуцію дії та атрибуцію відповідальності: можливість приписати системі конкретну дію розглядається як підстава для визнання її агентності, але не як достатня умова визнання її суб'єктом відповідальності. Як критерій переходу від інструментальності до агентності аналізується здатність системи самостійно визначати спосіб досягнення поставленої мети; особливим тестом такої агентності постає можливість атрибуції системі помилки. Обґрунтовується, що атрибуція помилки не тотожна атрибуції вини. Перехід від агента до суб'єкта відповідальності розглядається через суб'єктивний, нормативно-соціальний та темпоральний виміри. Окремо поставлено проблему безперервності тотожності штучної системи в умовах можливості її копіювання та відновлення попереднього стану.

Ключові слова: штучний інтелект, відповідальність, атрибуція відповідальності, атрибуція дії, агент дії, агентність, суб'єкт відповідальності, помилка, вина.

The paper examines the problem of attributing responsibility to artificial intelligence by distinguishing three statuses of an artificial system: instrument, agent, and subject of responsibility. A distinction is proposed between the attribution of action and the attribution of responsibility: the possibility of attributing a particular action to a system is regarded as a basis for recognizing its agency, but not as a sufficient condition for recognizing it as a subject of responsibility. The system's capacity to determine autonomously the means of achieving a given goal is examined as a criterion for the transition from instrumentality to agency, while the possibility of attributing an error to the system is considered a specific test of such agency. It is argued that the attribution of error is not identical to the attribution of culpability. The transition from agent to subject of responsibility is examined through the subjective, normative-social, and temporal dimensions of responsibility. Particular attention is paid to the problem of continuity of identity in artificial systems in view of their potential capacity to be copied or restored to a previous state.

Keywords: artificial intelligence, responsibility, attribution of responsibility, attribution of action, agent of action, agency, subject of responsibility, error, culpability.

Вступ

Проблема відповідальності традиційно пов'язана з проблемою дії та її суб'єкта. Щоб поставити питання про відповідальність за певний вчинок, необхідно спочатку встановити, кому може бути атрибутована сама дія. У праці «Намір» (*Intention*) Г. Е. М. Енскомб пов'язує дію з поняттям наміру та особливим способом, у який суб'єкт знає, що саме він робить [1]. Такий підхід передбачає розмежування між суб'єктом дії та засобом її здійснення: інструмент може розширювати можливості людини, однак сам не є автором вчинку.

Розвиток автономних систем штучного інтелекту ускладнює цю модель. Система може не лише виконувати безпосередньо визначену людиною операцію, а й обирати способи досягнення поставленої мети, реагувати на зміни середовища та формувати результати, які не були наперед визначені розробником або користувачем. А. Маттіас описує один із наслідків цієї ситуації як «розрив відповідальності» (*responsibility gap*): традиційне приписування відповідальності оператору або розробнику стає проблематичним, коли поведінка системи вже не може бути ними повністю передбачена [3].

Однак непередбачуваність системи не означає, що відповідальність автоматично може бути перенесена на неї. Необхідно розмежувати два питання: кому атрибутується дія і кому атрибутується відповідальність за неї. Таке розмежування простежується у концепції Л. Флоріді та Дж. В. Сандерса, яка дозволяє розглядати агентність без обов'язкового приписування агенту всього комплексу властивостей людської моральної суб'єктності [4].

Для аналізу штучного інтелекту видається недостатньою бінарна опозиція між інструментом і суб'єктом. Між ними доцільно виокремити третю категорію — агента. Інструмент є тим, за допомогою чого діє суб'єкт; агент — тим, кому може бути атрибутована дія; суб'єкт відповідальності — тим, кому може бути атрибутована не лише дія, а й відповідальність за неї.

Метою роботи є визначення концептуальних меж переходу між цими трьома статусами. Вихідним припущенням є те, що атрибуція дії та атрибуція відповідальності не є тотожними: можливість приписати системі певну дію може бути підставою для визнання її агентності, однак ще не є достатньою підставою для визнання її суб'єктом відповідальності.

1. Штучний інтелект як інструмент

Найпростішою моделлю визначення статусу штучного інтелекту є його розуміння як інструмента. У її межах ШІ може розширювати можливості аналізу, автоматизувати операції та виконувати складні послідовності дій, однак джерелом мети та суб'єктом діяльності залишається людина.

Таке розуміння має тривалу історію в дискусіях про штучний інтелект. Дж. Серль у праці

«Розуми, мізки та програми» (*Minds, Brains, and Programs*) розрізняє «слабкий» і «сильний» ІІІ: у першому випадку комп'ютер є інструментом дослідження мислення, тоді як другий передбачає, що належним чином запрограмована система сама може розуміти та мати когнітивні стани [5]. Дж. Вайценбаум у праці «Комп'ютерна могутність і людський розум: від судження до обчислення» (*Computer Power and Human Reason: From Judgment to Calculation*) проводить принципову межу між обчисленням і людським судженням та застерігає проти перенесення на комп'ютер рішень, що належать до сфери людської відповідальності [6]. Г. Дрейфус, своєю чергою, критикує припущення, що людське розуміння може бути вичерпно відтворене через формальні правила та маніпуляцію символами, наголошуючи на ролі контексту, практичного залучення до світу та фонового знання [7].

Спільним для цих позицій є принципове застереження: технологічна складність сама по собі не перетворює систему на суб'єкта. Так само складність, непрозорість або непередбачуваність результату ще не є достатніми критеріями переходу від інструмента до агента.

Ключовим для інструментальної моделі є те, що конкретна дія інструмента функціонально включена в дію його користувача. Інструмент утворює причинну ланку між рішенням суб'єкта та результатом, але не стає окремим адресатом атрибуції дії. Тому значна частина застосувань ІІІ може залишатися в межах цієї моделі: вислів «ІІІ зробив» у такому випадку є лише мовним скороченням для складнішого причинного опису процесу.

Межа інструментальної моделі виявляється тоді, коли між поставленою людиною метою та конкретною дією системи виникає простір, у якому спосіб досягнення мети вже не визначається людиною безпосередньо. Система може аналізувати середовище, вибирати між альтернативами та коригувати поведінку залежно від результатів попередніх дій. Проте й тут сама непередбачуваність не є критерієм агентності: непередбачуваними можуть бути природні процеси або складні механізми, яким ми не атрибуємо дії.

Отже, перехід до агентності починається не там, де система стає достатньо складною, а там, де для опису події стає недостатнім твердження, що людина здійснила конкретну дію за допомогою системи. Питання «хто використовував інструмент?» поступається питанню «кому належить ця конкретна дія?».

2. Штучний інтелект як агент

Один із продуктивних підходів до агентності запропонували Л. Флоріді та Дж. В. Сандерс. Вони характеризують агента через три властивості: інтерактивність, автономність та адаптивність [4]. Такий підхід дозволяє говорити про агентність незалежно від попереднього вирішення питання про наявність у системи свідомості.

Для проблеми відповідальності, однак, функціональних характеристик недостатньо. Принциповим є питання: чи можемо ми атрибутувати системі конкретну дію? Агентність у такому разі постає не лише як сукупність властивостей системи, а як відношення між системою, дією та її

результатом.

Це відношення проявляється тоді, коли система отримує не конкретну послідовність команд, а мету, в межах якої самостійно визначає спосіб її досягнення. Людина залишається джерелом початкової мети, проте вже не є безпосереднім автором кожного рішення. Виникає відмінність між твердженнями «людина здійснила дію за допомогою системи» і «людина поставила мету, а система здійснила дію для її досягнення».

Зовнішнє походження мети саме по собі не виключає агентності. Людський агент також діє в межах поставлених іншими завдань, соціальних норм і фізичних обмежень. Визначальним є те, чи задає зовнішній суб'єкт конкретний спосіб реалізації мети, чи цей спосіб формується системою в процесі взаємодії із середовищем.

Особливу евристичну цінність тут має феномен помилки. Поки система функціонує відповідно до очікувань, межа між інструментом і агентом може залишатися непомітною. Помилковий результат змушує поставити питання: хто саме помилився?

Не кожна несправність є помилкою в цьому значенні. Якщо неправильний результат безпосередньо спричинений технічною відмовою або помилковою командою користувача, немає підстав переносити авторство на систему. Інша ситуація виникає тоді, коли система коректно отримала мету й необхідну інформацію, але в процесі самостійного вибору способу дії сформувала хибне рішення. У такому випадку поняття «помилка системи» перестає бути лише метафоричним описом несправності.

Тому помилка може розглядатися як тест на атрибуцію дії. Якщо конкретна помилкова дія виникає в процесі автономного вибору системою способу досягнення мети і не може бути безпосередньо ототожнена з конкретним людським рішенням, виникають підстави атрибутувати цю дію самій системі.

Проте тут проходить принципова межа: атрибуція помилки не тотожна атрибуції вини. Вислів «система помилилася» ще не означає, що система винна. Перше твердження стосується авторства дії; друге передбачає нормативну оцінку агента та можливість атрибутувати йому відповідальність за цю дію.

Отже, агент займає проміжне положення між інструментом і суб'єктом. На відміну від інструмента, його діяльність не може бути повністю редукована до конкретної дії користувача; на відміну від суб'єкта, саме авторство дії ще не є достатньою підставою для моральної відповідальності. Агент — це той, кому може бути атрибутована дія, але кому ще не обов'язково може бути атрибутована відповідальність за неї.

1. Від агента до суб'єкта відповідальності

Перехід від агента до суб'єкта відповідальності потребує додаткових підстав. Визнання агентності дозволяє відповісти на питання «хто це зробив?», але не обов'язково на питання «хто

повинен за це відповідати?».

Одну з меж дозволяє провести концепція раціональної дії Дж. Серля. У праці «Раціональність у дії» (*Rationality in Action*) він виділяє три «розриви» (*gaps*) у структурі свідомої вольової дії: між підставами та рішенням, між рішенням і початком дії та між початком тривалої дії і її продовженням [2]. Ці розриви не означають просто непередбачуваності поведінки: вони існують для самого суб'єкта як переживання того, що попередні причини не визначають його наступної дії як каузально необхідної. Тому функціональний вибір штучної системи ще не свідчить про існування суб'єкта, який переживає можливість вибору.

Ця межа пов'язана зі сформульованою Д. Чалмерзом «важкою проблемою свідомості»: пояснення когнітивних функцій саме по собі не пояснює, чому їх виконання супроводжується суб'єктивним досвідом [8]. Навіть система, яка функціонально оцінює альтернативи, виправляє помилки або пояснює власні рішення, не обов'язково є суб'єктом, для якого цей вибір, помилка чи пояснення мають значення.

Однак самого внутрішнього досвіду також недостатньо для пояснення відповідальності. Вона виникає у відносинах з іншими суб'єктами, де існують правила, очікування, схвалення, осуд і вимога пояснення власної дії. Тут важливою є «інтенціональна настанова» (*intentional stance*) Д. Деннета: поведінку системи можна пояснювати через приписування їй переконань, бажань та раціональних цілей, якщо така стратегія продуктивна для її пояснення та прогнозування [9].

Проте пояснювальна атрибуція ще не є нормативною. Ми можемо говорити, що система «хоче», «вирішує» або «помиляється», не визнаючи її суб'єктом відповідальності. Такий перехід відбувається тоді, коли агент включається у спільний нормативний простір: його дія може бути поставлена йому у вину, від нього можна вимагати відповіді, а ця відповідь має значення для його відносин з іншими.

Саме поняття «відповідати» виявляє реляційний характер відповідальності. Відповідати — означає не лише бути причиною наслідку або автором дії, а бути тим, до кого інший може звернути нормативну вимогу щодо цієї дії. Здатність системи згенерувати пояснення ще не означає відповідати у цьому сенсі: необхідно, щоб вимога стосувалася її самої як автора дії та мала значення для її подальшого становища у спільному нормативному просторі.

Звідси постає ще одна характеристика суб'єкта відповідальності — скінченність і безперервність його існування. Людський суб'єкт не може необмежено повернутися до попереднього стану або створити ідентичну копію самого себе. Його вчинки включаються в незворотну біографію, змінюючи як його самого, так і його відносини з іншими. Скінченність сама по собі не створює моральної суб'єктності, однак у поєднанні із суб'єктивністю та соціальністю надає відповідальності особливої ваги: дія незворотно входить в історію конкретного суб'єкта та спільний світ інших.

Щодо штучного інтелекту це породжує специфічну проблему. Штучна система потенційно

може бути скопійована, відновлена з попереднього стану або запущена одночасно в кількох екземплярах. Якщо один із двох ідентичних до певного моменту екземплярів здійснює дію, що породжує відповідальність, хто саме повинен відповідати: цей екземпляр, усі його копії чи система як абстрактна тотожність? Можливість «відкотити» систему до стану, що передував дії, так само ставить питання про безперервність суб'єкта відповідальності.

Це не доводить неможливості штучної суб'єктності, але показує недостатність самої функціональної агентності. Необхідно встановити щонайменше три умови: чи існує суб'єкт, для якого власна дія має значення; чи включений він у нормативні відносини з іншими; чи існує достатня безперервність його тотожності, щоб здійснена дія залишалася саме його дією в часі.

Отже, суб'єкт відповідальності може бути визначений як агент, якому не лише атрибутується дія, але який здатний перебувати у нормативному відношенні до цієї дії, до самого себе та до інших суб'єктів у часі. Відповідальність виникає на перетині цих відносин: дія повинна належати агенту, мати значення для самого суб'єкта, бути предметом вимоги з боку інших і залишатися його дією в межах безперервності його існування.

Висновки

Запропоноване розрізнення інструмента, агента та суб'єкта дозволяє розвести три рівні атрибуції. Інструмент є тим, за допомогою чого діє інший суб'єкт; агент — тим, кому може бути атрибутовано конкретну дію; суб'єкт відповідальності — агентом, якому може бути атрибутовано також і відповідальність за неї. Перехід між цими статусами не є автоматичним: складність інструмента не робить його агентом, а автономність агента — суб'єктом відповідальності.

Межа між інструментом і агентом виникає тоді, коли конкретний спосіб досягнення поставленої мети формується самою системою і не може бути безпосередньо редукований до конкретного людського рішення. Особливим тестом цієї межі є помилка: можливість атрибутувати системі помилкову дію може бути підставою для визнання її агентності, але атрибуція помилки не тотожна атрибуції вини.

Перехід від агента до суб'єкта потребує іншого рівня обґрунтування. Відповідальність передбачає суб'єкта, для якого власна дія має значення, його входження у нормативні відносини з іншими та достатню безперервність його тотожності в часі. Можливість копіювання або відновлення штучної системи особливо гостро демонструє проблему такої безперервності.

Отже, питання про відповідальність штучного інтелекту доцільно формулювати не як бінарну альтернативу між інструментом і суб'єктом, а як послідовність трьох питань: через що здійснено дію, кому належить сама дія і хто може за неї відповідати? Перше визначає інструмент, друге — агента дії, третє — суб'єкта відповідальності. Таке розмежування дозволяє визнавати можливу агентність штучного інтелекту, не редукуючи його до пасивного засобу, але й не наділяючи його передчасно статусом морального суб'єкта.

Література

1. Anscombe G. E. M. *Intention* / G. E. M. Anscombe. — 2nd ed. — Cambridge, MA: Harvard University Press, 2000. — IX, 94 p. — ISBN 0-674-00399-3.
2. Searle J. R. *Rationality in Action* / John R. Searle. — Cambridge, MA: The MIT Press, 2001. — XVI, 303 p.
3. Matthias A. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata / Andreas Matthias // *Ethics and Information Technology*. — 2004. — Vol. 6, No. 3. — P. 175–183. — DOI: 10.1007/s10676-004-3422-1.
4. Floridi L. On the Morality of Artificial Agents / Luciano Floridi, J. W. Sanders // *Minds and Machines*. — 2004. — Vol. 14, No. 3. — P. 349–379. — DOI: 10.1023/B:MIND.0000035461.63578.9d.
5. Searle J. R. *Minds, Brains, and Programs* / John R. Searle // *Behavioral and Brain Sciences*. — 1980. — Vol. 3, No. 3. — P. 417–424. — DOI: 10.1017/S0140525X00005756.
6. Weizenbaum J. *Computer Power and Human Reason: From Judgment to Calculation* / Joseph Weizenbaum. — San Francisco: W. H. Freeman, 1976. — 300 p.
7. Dreyfus H. L. *What Computers Still Can't Do: A Critique of Artificial Reason* / Hubert L. Dreyfus. — Cambridge, MA: The MIT Press, 1992. — 408 p.
8. Chalmers D. J. Facing Up to the Problem of Consciousness / David J. Chalmers // *Journal of Consciousness Studies*. — 1995. — Vol. 2, No. 3. — P. 200–219. — URL: <https://consc.net/papers/facing.pdf>.
9. Dennett D. C. *The Intentional Stance* / Daniel C. Dennett. — Cambridge, MA: The MIT Press, 1987. — 400 p.