УДК 141+001.18

© Ігор Снегірьов
Чернівецький національний університет імені Юрія Федьковича
snegilecter@rambler.ru

ШТУЧНИЙ ІНТЕЛЕКТ ЯК ЧИННИК НЕЛІНІЙНОГО РОЗВИТКУ СОЦІАЛЬНИХ СИСТЕМ

З позицій методології нелінійного прогнозування розкриваються перспективи і ризики виникнення штучного інтелекту. Виявляються причини оптимістичних прогнозів, а також загрози в «горизонтах» сингулярності, зумовленої створенням машинного розуму. Особлива увага приділяється концепції дружньої розумної машини в умовах інтелектуального вибуху. Здійснено спробу філософської рефлексії декількох варіантів згубної відмови, а саме: порочної реалізації та інфраструктурної надмірності. *Ключові слова: штучний інтелект, сингулярність, нелінійність, антропоморфізація, «бог у ящику», тест Тьюрінга, три закони робототехніки, «чорний ящик», генетичні алгоритми.*

Одним з основних чинників, які впливають на відносно стійкий розвиток людини і соціальних систем, постає діяльність, яка в міру ускладнення форм, засобів і способів інновацій поступово перетворювалася на технології. Їх негентропійний потенціал і технічна реалізація в різних сферах життєдіяльності з древніх часів впливали на формування наукової раціональності сучасного типу, що, безсумнівно, корелює з виникненням інформаційного суспільства з характерним для нього посиленням ролі статистичних закономірностей і чинника нелінійності. З середини XX століття наука відіграє провідну роль у системі суспільного виробництва, а наукомісткі технології стали претендувати на роль нового атрактора, що детермінує нові гносеологічні вектори й аксіологічні горизонти розвитку соціальних систем. Але чим складніша соціальна система, тим більше вона підвладна впливу стохастичних чинників, які створюють нелінійний простір для подальших шляхів розвитку, що формуються в результаті актуалізації фазових переходів – вимушених відповідей нерівновагої структури на загрозу зниження стійкості.

У цьому контексті прогрес набуває характеру не самоцілі, не самозумовленої цінності, а постає в ролі способу збереження складної цілісності. Шляхи досягнення такого стану нелінійні, тому що заздалегідь прорахувати їхню кількість, ступінь детермінації та небезпеки на певний період часу неможливо. Майбутнє може стати як і «кращим» від сьогодення за строго визначеними параметрами, так і «гіршим» за іншими параметрами. І технології, особливо наукомісткі, які можуть стати причиною екзистенціальної кризи глобальних масштабів, віграють у цьому не останню роль. Розв'язання одних суперечностей «запускає» нелінійний ланцюг багатьох інших, нових, ще більш неоднозначних проблем. Надалі це зумовлює виникнення векторів еволюційних змін: від більш стохастичних («природних») до менш імовірних станів. Відповідно до нелінійної моделі прогрес як «відокремлення від природної ніші» означає відновлення відносної стійкості системи на більш високому рівні нерівноваги.

У світлі сказаного слід підкреслити, що «технологізація» суспільства і прогрес становлять нерозривну єдність, але, на жаль, не завжди органічну, як би того хотілося суб'єктові історичного процесу. Сьогодні вчені та філософи визнають, що саме наукомісткі технології – це той чинник і атрактор, які детермінують розвиток соціальних систем, а наука давно перетворилася на могутню продуктивну силу, яка час від часу, ігноруючи гуманістичні ідеали, заганяє себе в цинічну сферу, позначену формулою «мета виправдовує засоби».

Ілюстрацією подібного сценарію може слугувати класичний роман М. Шеллі «Франкенштейн, або Сучасний Прометей» [14], лейтмотивом якого є ідея, що наука, як і науковий метод, без аксіологічних засад і морально-етичних мотивацій ризикує перетворитися на знаряддя м'ясника, який намагається прикрити високими цілями висохлий кістяк особистих амбіцій і прагнень. Будь-яка наукомістка технологія спроможна як запустити нелінійний простір нових ризиків, так і стати потенційним джерелом загрози. І в цьому розумінні, на наш погляд, особливо цікавою стає проблема штучного інтелекту.

Створення думаючої машини, як і вона сама, є процесом із загостреним режимом протікання, однією з відмітних сторін якого є розростання мікрофлуктуацій, внаслідок чого «миша народжує гору». В нелінійній і нестійкій соціальній системі в планетарному масштабі це може мати катастрофічні наслідки, всупереч численним думкам апологетів цієї технології, тому що

будь-яку технічну інновацію можна використати як на зло, так і на благо. Історія вчить, що, як правило, перше є закономірним наслідком другого, або ж ці іпостасі постають двома сторонами однієї медалі. Ризики і ступінь відповідальності виростають у сотні разів, коли мова заходить про технології, дітище якої буде мати можливість удосконалювати саме себе. А одним із варіантів реалізації сценарію творення машинного розуму може стати інтелектуальний вибух, внаслідок якого надрозумний агент, який вирвався у світову мережу почне стрімко експоненціально навчатись і впродовж найкоротшого інтервалу часу вийде з-під контролю недалекоглядного людства та, можливо, ідентифікувавши вид homo sapiens як конкуруючий, використає його як ресурс, або просто знищить.

Але перш ніж аналізувати можливі негативні флуктуації, що можуть виникнути як у процесі, так і в результаті виникнення штучного інтелекту, проаналізуємо причини, внаслідок яких, незважаючи на очевидні ризики, людство все-таки жадає появи в цьому світі універсального всемогутнього раба, який легко може перетворитися на штучного бога. Чому ж багато представників наукового співтовариства не помічають або ж ігнорують можливість такої трансформації?

Насамперед, це помилка завищеного оптимізму, зумовленого багатьма чинниками, спектр яких досить широкий – від психологічних до економічних. Адже штучний розум, на думку багатьох, здійснить, або значно прискорить якісний переворот у багатьох наукових і технологічних галузях. Безумовно, що поява багатьох наукових відкриттів може виявитися неможливою завдяки тільки одному типу наукомістких технологій, будь то нанотехнології чи розумна машина. Як і не можливий розвиток у межах однієї наукової галузі. Така ситуація зумовлена кореляційним характером наукових революцій, значна частина яких розпочалася ще в минулому столітті у сфері інформаційних, комунікаційних і біологічних технологій. Стрибкоподібний розвиток когнітивних наук в останні десятиліття дозволив говорити експертам про наближення нової наукової революції. Особливий інтерес і значення має кореляційне взаємопроникнення саме інформаційних, біологічних, нано і когнітивних технологій, що дістало назву NBIC-конвергенція.

Штучний інтелект відіграє в такій синергетичній системі одну з провідних ролей. Наприклад, штучний суперінтелект у перспективі може створити самовідтворюювані машини молекулярної збірки, що дістали назву наноассемблери, давши обіцянку людству, що їхнє використання принесе винятково користь. Об’єктивно ж надалі суперінтелект замість того, щоб перетворювати пісок на золото, почне перетворювати матеріали на програмовану матерію, яку потім він зможе перетворювати на що завгодно – від комп’ютерних процесорів до космічних мегамостів для колонізації Всесвіту. Але все це виглядає поки що на рівні сюжету для художнього твору.

Стає зрозуміло, що людство у марнолюбній гонитві за лавровою гілкою першості мало перед чим зупиниться, тим більше що далеко не останню роль у такій гонитві відіграє фінансова сторона справи. Адже ні для кого не секрет, що одним з основних складників продуктивних сил є прибуток. А штучний інтелект, можливо, не тільки прискорить цей процес, а й сам стане новим видом способу виробництва. Питання лише в тому, чи буде в цей процес включена сама людина?

Безкрайній оптимізм підживлює також і комплекс складних психологічних причин, сукупність яких дістала назву “антропоморфізм” – екстраполяція людських якостей на можливості надрозумного агента. «Щоб розмова про надразум вийшла осмисленою, потрібно заздалегідь усвідомити, що надрозум – це не просто ще одне технічне досягнення, ще одне знаряддя, що збільшить людські можливості. Надрозум – щось принципово інше. Цей момент потрібно всіляко підкреслювати, оскільки антропоморфізація надрозуму – плідний ґрунт для оман» [4, с. 224].

Штучний інтелект – якісно нове в технологічному розумінні, за словами Н. Бострома, в силу того, що його виникнення змінить суть прогресу, виключивши із цього процесу людину. Машинний розум буде ні на що не схожий. Будучи результатом людської діяльності, він буде прагнути саморозвитку за одним, лише йому відомим сценарієм, який не залежить від людини. У нього не буде особистісних мотивів, тому що не буде людської сутності.

Отже, антропоморфізація машини є джерелом помилкових уявлень про те, що можна створювати безпечні машини, наділені розумом, і що це рано чи пізно не приведе до катастрофи. Почасти така віра в «дружелюбних роботів» заснована на трьох законах робототехніки, які чомусь широка публіка сприймає як даність і сліпо покладається на них. Тоді як насправді ці три закони в

розповіді «Хоровод» [2, с. 122-164] були запропоновані фантастом Айзеком Азімовим. За сюжетом, вони жорстко пов'язані в нейронні мережі «позитронного» мозку роботів:

1. Робот не може заподіяти шкоди людині чи своєю бездіяльністю допустити, щоб людині була заподіяна шкода.
2. Робот повинен коритися командам людини, якщо ці команди не суперечать Першому законові.
3. Робот повинен піклуватися про свою безпеку доти, доки це не суперечить Першому і Другому законам.

По суті, ці закони є своєрідною адаптованою інтерпретацією заповіді «Не убий», християнського розуміння того, що до гріха можна прийти як роблячи вчинки, так і «сидячи на березі ріки», клятви Гіппократа. Важливе інше: у літературній спадщині А. Азімова ці закони так чи інакше дають збій. Та й здебільшого творчість відомого фантаста спрямована на те, щоб продемонструвати, наскільки самовпевнена людина у своїх спробах контролювати сутності, які перевершують її практично за всіма параметрами. У вищезгаданій розповіді «Хоровод» геологічна експедиція на Марсі довірила роботові перевезення токсичної для нього речовини. Але місія виявляється невиконаною в силу того, що робот потрапляє в пастку зворотних зв'язків другого і третього законів. Він би так і залишився в цьому лабіринті сумнівів, якби не прийшли йому на допомогу геологи. Розповіді А. Азімова буяють сюжетами, в яких досягти мети можна лише в обхід трьох законів (тобто первісно висувається постулат про сумнівний характер цих «трьох заповідей).

Оригінальні й захоплюючі сюжети в художній літературі та безпека в реальному світі – речі різні. Різноманіття та складна нелінійність реального світу більш непередбачувані та стохастичні, ніж будь-яка вигадка генія. Випадковість у літературі фіксована й ув'язнена в межах цього семантичного поля (оповідання, повісті, роману). Випадковість же в реальному світі, з одного боку, «виношує» наступна мить, а з іншого, попередньою миттю може бути «виношена» й розчинена в потоці подій, не піддається фіксації та контролю до реалізації з потенції та її важко виявити й описати після. Особливо, коли ми маємо справу з об'єктами, виникнення та поведінка яких не піддається стандартним методам. До них, безсумнівно, належить і поява штучного інтелекту. Згаданих трьох законів недостатньо. Річ у їхньому недостатньо чіткому формулюванні та розпливчастій семантиці понять, якими ці закони оперують. Так, звична для нас сьогодні дихотомія між живим (людиною) і неживим (роботом), імовірно буде розмита, коли наука навчиться підсилювати тіло і мозок людини за допомогою, наприклад, комп'ютерного інтерфейсу. І що тоді буде називатися людиною? Аморфні й інші поняття законів: «шкода», «безпека», «команди».

Задля справедливості варто сказати, що А. Азімов пізніше доповнив свої три закони своєрідним «приквелом», нульовим законом, який забороняє роботам шкодити людству в глобальному масштабі. Але якісно це проблему не розв'язує. Ці закони є монументальним обеліском великому генієві автора художньої прози і самому жанру, але вони найбільш цитовані, коли здійснюються спроби вибудувати стратегію майбутнього співіснування виду homo sapiens і надрозумного агента. Виникає очевидне, але лякливе запитання: невже Три закони це все, що ми маємо сьогодні?

У той час, як «темні сторони» роботизації вже дають про себе знати і реальний стан справ змушує замислитись. У світі в понад п'ятдесяти країнах сьогодні ведуться розробки мілітаризованих думаючих машин. І вже є перші вияви неякісного програмування, що зумовили страшні наслідки: під час воєнних дій на Близькому Сході бойові дрони, оснащені автоматичною зброєю, після використаня проти своїх, були виведені з ладу. Збій у роботі програми роботизованої зенітної гармати спричинив загибель девяти і важкі поранення п'ятнадцяти солдатів 2007 р. на африканському континенті [див.: 13, с. 128]. Дуже симптоматичний той факт, що тривалість цього інциденту була меншою від секунди. Отже, виникає думка, що дискусії про етику і про технічні інновації відбуваються на різних планетах.

Концепція техно-гуманітарного балансу постулює потребу в наявності відповідного духовного рівня цивілізації для адекватної «притирання» до технологічних інновацій різного ступеня складності [див.: 8]. Технологія штучного розуму, як і розподіл ядер, як і багато інших, – технологія подвійного призначення. Розщеплення ядра є джерелом освітлення міст, вона живить цивілізацію, але також може її й спопелити. Трагедія Хіросіми і Нагасакі – дивовижний приклад того, як «технологічне насіння» впало в «незоране гуманітарне поле» цивілізації. В історії таких прикладів цілком достатньо. Якщо стрибок від штучного інтелекту людського рівня до штучного суперінтелекту відбудеться за сценарієм стрімкого інтелектуального вибуху, то людству ще раз доведеться зіткнутись із загрозою планетарного масштабу. Питання в тому – чи вистачить нам мудрості пережити черговий «виклик».

Друге перекручування в осмисленні ризиків виникло внаслідок популярності теми штучного розуму у світі розваг. Як правило, обговорення небезпек, які пов'язані зі штучним інтелектом, відбуваються у відриві від контексту, з підміною понять і не відзначаються професійною оцінкою та аналітичною глибиною. Зрозуміло, що в академічних наукових і філософських колах цю проблему не обходять стороною, але найчастіше їй не приділяють належної уваги. Більшість фахівців уважають створення штучного суперрозуму лише справою часу і перебувають у приємній ейфорії від тих перспектив, які вимальовуються на горизонті.

Щойно мова заходить про неприємні прогнози, багато представників ЗМІ (технічні журналісти, блогери, редактори) рефлексивно не приймають їх всерйоз. Така реакція детермінована звичайним небажанням дістатися до суті проблеми, що свідчить про неспроможність висунутих ними контраргументів. Для більшості обивателів, на яких розрахована продукція ЗМІ, проблема штучного розумного агента здається чимось далеким і абстрактним, а отже, вона не є для журналістів справою першорядною. Привабливішою для технічної журналістики є сфера розваг: плазми на квантових точках, більше ємнісні носії інформації та інші новинки програмного ринку. Кінематограф подарував нам десятки апокаліптичних фіналів, які затіяв бунтівний машинний інтелект («3 машини», «Космічна одісея 2000 року», «Термінатор», «Матриця», «Я-Робот» й інші), що породило ефект «комфортного абстрагування» від всерйоз навислої небезпеки. Ці «техномонстри» доставили глядачеві стільки приємних годин «безпечного лоскотання нервів», що зрештою в підсвідомості більшості закріпився ефект надуманої небезпеки. Адже, по суті, побачене на екрані – результат чиєїсь уяви і не має до реального світу жодного стосунку. Інакше кажучи, розтиражованість художніх інтерпретацій небезпек виникнення машинного розуму (кіно, література) зіграла роль своєрідного щеплення, після якої можливість серйозного осмислення й аналізу катастрофічних ризиків стала практично неможливою. Не викликає побоювань і створена на початку XXI ст. роботизована імітація відомого фантаста Филипа К. Діка. З роботизованою машиною можна обговорювати творчість автора. «Люб'язна імітація імітує люб'язність». Як у цьому контексті не згадати: «Найбільша хитрість диявола – змусити всіх повірити, що його не існує».

Когнітивне перекручування – ще одна причина помилкового уявлення, що виникнення розумної машини рідко має деструктивний характер для людини. І на цій проблемі ми зупинимося детальніше. Річ у тім, що людина в процесі своєї діяльності приймає рішення, керуючись, здебільшого, власним досвідом, який найчастіше не піддається рефлексії, а має, скоріше, стихійно-буденний рівень. Об'єктивно ж завжди залишається ймовірність того, що дослідник урахував не всі вхідні дані експерименту й флуктуація одного з неперелічених елементів вплине на кінцевий результат. А в результаті того, що будь-який досвід має обмежений характер, у гносеологічному аспекті ми неминуче зіштовхуємося з проблемою невизначеності омани.

Є ймовірність того, що відсутність контакту з об'єктом у ретроспективному аспекті унеможливлює побудови каузальних зв'язків між виникненням суперрозумного агента і зникненням цивілізації в глобальному масштабі. Сьогодні жодна людина не стикається з жодною серйозною подією, що має трагічні наслідки, до початку й до основи якої був би причетний штучний розумний агент. Штучний інтелект поки не розцінюється як джерело екзистенціальної загрози. Створюється думка, що для того, щоб людству оцінити всілякі ризики, причиною яких може бути штучний інтелект, йому потрібно опинитися на грані життя і смерті. Зіткнення з розумом, який перевершує наш, не буде мати нічого спільного з обмеженими в часі та просторі терористичними атаками, ядерними зимами й іншими катастрофами, які мають ендо-техногенний характер. У результаті актуалізації сценарію, що вирвався з неконтрольованого штучного суперрозуму, глобальна людська раса, очевидно, залишить після себе лише слід у вигляді казкових історій, які будуть розповідати роботи своїм дітям перед сном.

Звільнений з «ящика» штучний інтелект має ще одну принципово якісну відмінність від техногенних катастроф. Сьогодні людство зіштовхувалося лише з тими подіями, негативні наслідки яких були подолані. У випадку ж «повсталого демона в машині» має місце самовдосконалювана і самовідтворювана розумна програма, яка потенційно може існувати вічно. І зрозуміло, що в цієї усвідомлюючої себе системи, будуть базові потреби. Згідно з С. Омохундро, їх чотири: ефективність, самозахист, ресурси, творчість [9], що дублює потреби людини й історія вчить до чого, як правило, приводять спроби егоїстичного їх задоволення. Наскільки потенційно небезпечна кожна з них сьогодні сказати складно, але якщо загрозу від перших трьох ми можемо уявити

загалом, то потреба у творчості при першому осмисленні ніби й не несе жодних ризиків, а скоріше навпаки. Але саме здатність до творчості є однією з основних умов існування та функціювання думаючої машини. Саме вона буде зумовлювати можливість самоідентифікації машини й усвідомлення нею себе як Я-буття, а отже й особистісної автономії.

Тут і виникає істотна проблема: створити штучний інтелект людського рівня – означає наділити машину свідомістю, однією з основних характеристик якої є здатність до цілевизначення, створення чогось нового, що неможливе без відносної самостійності функціювання програмного забезпечення. Це неминуче приведе до відносної втрати контролю людини над своїм дітищем, а за умов безперервного самовдосконалення машини – до повної автономії функціювання розумного агента. Іншими словами, наявність творчості в машини, що усвідомлює себе, є її детермінуючою характеристикою, що водночас і є ключем до її свободи, і цей ключ, так чи інакше, ми вручимо їй самі. Чи спроможні ми протистояти загрозі, причина якої співвідноситься з нами в розумовому розвитку приблизно так само, як ми і дощовий хробак? І чи зможемо ми подолати наслідки катастрофи, яка один раз почалась, але триває нескінченно?

Також однією із причин некоректної оцінки штучного розуму як початку зворотного відліку існування цивілізації є той факт, що феномен штучного розумного агента одержав своє відображення в іншому неоднозначному явищі, що дістало назву “сингулярність”.

Поняття “сингулярність” широко використовується в наукових і філософських колах і має різну контекстуально-семантичне забарвлення. У нашій статті вживання цього поняття має значення, вкладене у нього Р. Курцвейлом, який розуміє під цим якусь міру переходу кількості в якість, в якому стрибок технічного прогресу принципово трансформує буття людини. Розум перестане бути виняткова людською прерогативою і, на думку Р. Курцвейла, стане більш комп’ютеризованим, що зробить його в рази могутнішим, ніж сьогодні. Автор цієї концепції налаштований оптимістично, вважаючи, що такий перехід викорінить із людського існування такі негативні явища як голод, хвороби, а, можливо, у перспективі людина здобуде й вічне життя [див.: 5, с. 136-147].

На думку Р. Курцвейла, штучний розумний агент вплине на принципово новий якісний перехід світової цивілізації, але, як уже було зазначено, це спричинить розвиток і суміжних галузей наукомістких технологій – нано [5, 47]. Прогнози фахівців указують на те, що інтелектуальний вибух і наступна поява штучного надрозуму неминуче спричинить різкий стрибок у розвитку нанотехнологій. Багато експертів уважають, що пріоритет у виникненні повинен належати саме штучний суперрозум в силу того, що нанотехнології – занадто стохастичний інструмент, контроль над яким може виявитися для людини неможливим. Стає зрозумілим, чому оптимізм у контексті сингулярності виходить саме зі сфери розробок нанотехнологій, а не штучного розуму. Інженерія на атомарному рівні, можливо в майбутньому, дасть людині шанс обдурити смерть.

Але поряд із позитивними прогнозами існує і «ложка дьогтю». Наприклад, наніти, здатні до самовідтворення, перетворять навколишній світ на так званий «сірий слиз». Сценарій «сірого слизу» – один з імовірних апокаліптичних, і ця проблема – «темна територія» простору нанотехнологій. Міркуючи про зловісну сторону нанотехнологій, багато хто випускають з уваги принципову непередбачуваність, а отже, загрозу, пов’язану зі створенням штучного всемогутнього помічника, якщо той буде прогресувати за сценарієм стрімкого самовдосконалення, в ході якого машини, які перевершують за всіма параметрами людину, вийдуть з-під контролю та знищать глобальну цивілізацію.

Тож у результаті аналізу дослідницького досвіду з проблеми оптимістичного ставлення до розумних машин, умовно можна виокремити два підходи. Теоретико-світоглядні горизонти першого варіанту задають дослідження в дусі праць Р. Курцвейла [5], в яких майбутнє антропної компоненти нашої планети мислиться як украй позитивне. Негативні флуктуації в магістральному перебігу подій у такому осмисленні придушувалися б оптимізмом.

Другий підхід заснований на працях у стилі Д. Стібела [10], який осмислює ці проблеми з погляду прагматичного практицизму. Його прихильники трактують світову мережу як дедалі більш ускладнюваний мозок з мільйонами зв’язків і гарний той ділок, який з оптимальним ефектом для себе зуміє лавірувати в просторі інтернет-тенденцій та отримувати відтіля максимум прибутку.

Більшість експертів, задіяних у сфері наукомістких технологій не аналізують більше скептичний третій варіант, суть якого полягає в тому, що фінальним етапом розробок розумних

агентів, а потім і машин, які перевершують за інтелектом людину, стане не гармонійне єднання штучного інтелекту з природним, а перетворення людини на сировину для «тріумфального поступу» нового суб'єкта.

Взаємодія «традиційного» розуму людини із суперрозумом машини тотожна з розширенням сфер впливу технологічної західної цивілізації на суспільства аграрного типу, які або асимілювали їх, або перетворили на свій ресурс. Як приклад, досить згадати наступні антагонізми і чим вони завершилися: Колумб – Тіано, Пісарро – Інки, європейці – американські індіанці.

Що далі – homo sapiens проти машинного суперінтелекту?

Цілком імовірно, що теоретики в галузі наукомістких технологій вже осмислили всі «темні сторони» штучного інтелекту, але, проаналізувавши всі «за» і «проти», дійшли висновку, що мета виправдує засоби. Або ж розуміють, що точку неповернення пройдено і приймають неминучість будь-якого результату, постулюючи неможливість щось змінити. Висунуто ідею, відповідно до якої людина зможе осмислити й відкрити способи захисту від прогресуючого машинного розуму лише в процесі інтеграції цього феномену в наше буття. Ця взаємодія буде відбуватися поступово й у людства буде шанс прищепити розумному агентові «алгоритми слухняності» та створити дружній суперінтелект [див.: 3, с. 75]. Співзвучне цій ідеї розуміння неможливості усвідомлення всіх істинних ризиків штучного розуму теперішнього зі звичними для нас патернами буття [див.: 5, с. 34]. Іншими словами, приборкання дикого мустанга на дикому Заході не приводить до розуміння специфіки управління гоночним автомобілем на гірській дорозі.

Отже, проблема зазначеного підходу полягає в тому, що якщо загроза й визнається серйозною, то випливає вона з непередбачуваності «темного простору» інтелектуального вибуху й суперінтелекту, тоді як ризики і загрози від проміжних етапів створення штучного інтелекту аналізуються недостатньо серйозно, або ж зовсім не беруться до уваги. Інакше кажучи, грізна левиця, звичайно, є джерелом небезпеки для туриста в савані, але не можна забувати про потенційну загрозу, що йде від її милого левеняти. Концептуальні побудови градуалістів тією чи тією мірою конституюють поступовий характер стрибка від штучного інтелекту рівня людини до штучного суперрозуму, часовий інтервал якого може розтягтися від декількох років до десятиліть. Цей прогноз дозволяє розраховувати людині на період дружнього симбіозу з розумними машинами людського рівня й уможливлює сценарій, в межах якого людина зможе створити діючі важелі контролю над формованим суперінтелектом.

Але існує когорта дослідників, на думку яких людство позбавлене якого-небудь часового запасу. Річ у тім, що стрибок від штучного розуму людського рівня до штучного супермозку через самовдосконалення може відбутися різко. Відповідно до цього сценарію, стрімка трансформація онтологічних засад людства стане відображенням швидкого перетворення штучного розуму рівня людини на штучну надрозумну систему. Цей період може зайняти тижні, дні, а можливо й години. Цей сценарій дістав назву Busty Child.

Система, що складається з декількох суперінтелектуальних систем, кожна з яких у кілька тисяч разів буде перевершувати найрозумнішу людину, без особливих труднощів зможе подолати всі перешкоди, створені нами. Це можна порівняти з безкрайнім простором чужого інтелекту й однією його піщинкою. Для того, щоб загалом описати ті відчуття, які переживає людина в процесі взаємодії всього-на-всього з окремою програмою (Deep Blue – комп'ютерний шахіст фірми IBM), а не із самовдосконалюваною групою штучного суперрозуму, наведемо тут висловлення двох гросмейстерів, досить подібні за змістом, суть яких зводиться до наступного: «Начебто стіна на тебе насувається» [6, с. 48].

Досить показовий у цьому розумінні, проведений на початку XXI століття в Силіконовій долині експеримент, суть якого полягала в наступному: фахівець із вивчення штучного інтелекту уклав досить специфічну суперечку – хто переможе в грі, яку він назвав «штучний інтелект у ящику». Ставки на гру були досить високими. У ході цього експерименту роль машини зіграв ініціатор експерименту Е. Юдковскі [16], в ролі ланцюгових псів виступали мільйонери, які заробили свої статки на різних інтернет-проектах. Кожен з них, по черзі, виконував роль творця розумного агента, перед яким стояла мета не випустити з «ящика» штучний розум. «В'язень» і «Страж» підтримували контакт через онлайн-чат. Гра тривала не більше двох годин і всього їх відбулося п'ять. Можливість втомити «Вартового» мовчанням була передбачена, але не використовувалась. У результаті машина здобула перемогу в трьох сеансах. Як «Пандорі» вдалось

утекти невідомо, тому що однією з умов проведення експерименту була повна конфіденційність змісту кореспонденції між «В'язнем» і «Стражем» [16].

Цей експеримент доводить наші побоювання – якщо звичайний смертний зміг за допомогою слів «випустити з ящика Пандору», то ув'язнений суперінтелект, рівень якого буде в невідому кількість разів перевершувати людський, зробить це швидко і гарантовано. До того ж для машини досить усього лише раз вчинити втечу. І буде краще для всіх нас, якщо він виявиться дружнім.

Цікавий цей експеримент ще й тим, що, по суті, є варіантом тесту А. Тьюрінга, який розробив його в 50-му році минулого століття для визначення рівня інтелекту в машині. У процесі цього тесту програмі й опонентові, в ролі якого постає людина, задаються письмові запитання. Арбітр повинен зуміти визначити за відповідями, хто – людина, а хто – машина. Якщо ідентифікація неможлива, то комп'ютер здобуває перемогу. Виникає цілком закономірне запитання: як відрізнити оригінальний людський розум і мислення від достовірних спроб імітувати його? Де ця грань між об'єктом і образом об'єкта в мисленні суб'єкта? Інакше кажучи, машині зовсім не обов'язково думати, як людина, щоб давати відповіді, які ідентифікують її як таку. Машині досить зімітувати розумовий акт, даючи «антропоморфні» відповіді. Це стало «грою в імітацію» А. Тьюрінга, який вважав, що машини здатні до діяльності, яку суб'єкт спостереження легко прийме за розумну [див.: 12, с. 63]. По суті ж цей алгоритм у виконанні машини не має нічого спільного з розумовим актом, який здійснює людина. Тим самим, А. Тьюрінг вступає в полеміку з Д. Сьорлем, який вважав, що якщо машина не думає подібно до людини, то вона не розумна. Але більшість дослідників у галузі створення штучного розуму виражають солідарність із першим: якщо розумний агент здійснює розумні дії, імітуючи свідомість людини, то яка різниця, які процеси запускають його програми?

Тривал час пройти тест А. Тьюрінга було нерозв'язним завданням для машинного інтелекту й дослідників, які його створюють. Але влітку 2016 року в багатьох інтернет виданнях з'явилась інформація про те, що вперше комп'ютерній програмі, розробленої російськими авторами, вдалося це зробити. Ця програма видавала себе за 13 літнього хлопчика Євгенія Густмана й увела в оману 33% опонентів, спеціально запрошених із британського Університету Редінга. Один із членів команди, який розробляв програму, розповів, що головною ідеєю було переконати комісію в нібито широкому кругозорі, але який обмежений віком імітації [див.: 11].

Ця новина ще раз підтверджує той факт, що створення штучного розумного агента – справа недалекої перспективи і хитромудрих пасток у його запасі буде більш, ніж достатньо, а здатність імітувати людські якості, на додачу до всього сказаного, відображає вже цілком реальну можливість досягати своїх цілей найрізноманітнішими шляхами.

Чи можливо за наявності накопиченого за час розвитку науки прогностичного потенціалу передбачити ці цілі, способи і засоби, які машина для себе обере; а також рівень загрози, що випливає з перспектив появи штучного мислячого розуму? Оцінюючи можливість появи штучного розуму людського рівня, а потім і надрозумного агента з позиції однозначної позитивної рефлексії, дослідники тим самим звужують свій методологічний інструментарій лінійного, жорстко детермінованого прогнозування. Наявність у постнекласичній науці патернів лінійності зумовлена орієнтацією класики і некласики спиратися на динамічні закони, які описують світ з позицій жорсткої детермінації закритих систем. У розвитку таких систем не враховувалися флуктуаційні впливи середовища, які привносять у їхню впорядковану структуру фактор випадковості, відносної нестійкості. Прогноз на майбутню поведінку подібного типу систем був відносно простий – вивчивши і знаючи причину, ми однозначно можемо передбачити і наслідок.

Істотний прорив у методології системного прогнозування було здійснено в межах нелінійної методології, зініційованої теорією самоорганізації складних систем. Залежно від того, наскільки відкриті системи різної природи здатні перетворювати внутрішні неоднорідності своєї структури на корисний для себе потенціал, зберігаючи і збільшуючи рівень організованої складності та відносної стійкості, дослідникам удалося систематизувати специфіку протікання режимів із загостренням, а також способи ліквідації флуктуаційних відхилень, які ведуть до кризових станів. Як уже було сказано, процес створення машинного розуму, як і факт його наявності в майбутньому, належать до типу таких режимів, у яких випадкове мікровідхилення може спричинити низку непередбачуваних макронаслідків. Аналізуючи перспективи виникнення та розвитку штучного розуму з позицій нелінійного підходу, треба внести у прогностичну модель дані про специфіку штучної системи, тенденції розвитку, рівні загрози, а також вектори нелінійності. Прогностичне моделювання з

неврахованим вектором нелінійності не відображає всієї складності процесу і є лінійним. З подібними недоліками ми маємо справу, зіштовхуючись із прогностичними моделями, які пов'язані з технологією штучного інтелекту. У зв'язку із цим не варто недооцінювати творчий потенціал еволюції, стохастичний вплив якої не є чимсь постійним і стійким. І з цієї позиції, процес самовдосконалення штучного розуму, напевне, буде так чи інакше відповідати критеріям еволюційного розвитку. А. Азімов у своїх працях висував тезу, що когнітивні моделі можуть бути тотожні з компонентами психіки, іншими словами, він припустив, що роботи зможуть еволюціонувати. Його ідея про «примари в машині» нині набуває нового смислового наповнення. Він говорить про можливості виникнення випадково сформованих протоколів, які в перспективі можуть розвинутися в те, що ми називаємо поведінкою, а непередбачені вільні радикали можуть покласти початок креативності, свободі вибору, а можливо й душі [див.: 1, с. 253].

Також варто враховувати, що виникнення штучного суперінтелекту належить до розряду тих подій, наслідки від яких можуть бути глобальними, а ймовірність виникнення низька в силу того, що нічого подібного просто не відбувалося й емпіричний компонент людства тут дорівнює нулю. Прийти до однозначного рішення, маючи на озброєнні традиційні статистичні методи, завдання не з простих.

У процесі розробки штучно розуму варто мати на увазі, що фахівці не можуть собі дозволити права на помилку. В умовах процесу із загостреним режимом протікання це може мати наслідки «ефекту метелика» й у результаті ми одержимо щось зовсім чуже. Це можна порівняти з діями ведмежатника зі злому суперскладного сейфа під сигналізацією в банку. Якщо він із двадцяти цифр коду натисне правильно дев'ятнадцять, то двері не відкриються на п'ять відсотків, вони так і залишаться замкненими. Завиє серена і все закінчиться жалюгідно. При створенні штучного інтелекту помилка навіть в одну соту відсотка спричинить на сто відсотків непередбачуваний результат. Він не буде майже гарним, він буде цілком поганим.

Більшість звичних для людини технологій потенційно для неї небезпечні, як і будь-яка інновація – «ціпок з двома кінцями», про що вже йшлося. Очевидно, штучний інтелект – не виняток. Він не буде виявляти до вас негативу, але вашим атомам він може знайти інше застосування. Створюючи штучний розум із благими намірами й екстраполюючи ці позитивні емоції на машину, фахівці помилково вважають, що це є гарантом появи дружніх характеристик. Така недалекоглядність пов'язана з уже згаданим демоном антропоморфізму та лінійним детермінізмом, який не враховує флуктуацій різного ступеня важливості.

Вид homo sapiens виник у процесі самоорганізації матерії, що передбачає боротьбу за вільну енергію, інформацію та речовину. І, якщо на початку свого становлення в бутті людини домінували процеси стохастичного характеру, то з розвитком суспільства стихійні процеси були відсунуті на периферію завдяки свідомому чинникові та соціальним інститутам. Біфуркаційні фази долалися завдяки перемозі техно-гуманітарного балансу й усвідомленню суті та причин проблеми. Оглядаючись назад і аналізуючи сутність більшості антропологічних криз, можна дійти висновку, що їхня природа полягає в недооціненому ступені загрози наслідків, або ж у нерозумінні суті причин технологічного зльоту. Розвиток штучного інтелекту, напевне, теж буде передбачати його активне включення в процес природного добору, в боротьбу за енергію, речовину та інформацію. Та й, загалом, за все те, що вважатиме ресурсом.

Викликає, м'яко кажучи, побоювання той факт, що багато дослідників не зовсім розуміють, як працює вся система в цілому. І в цьому контексті явно недостатньо створювати розумну машину з чистим серцем і благими намірами, сподіваючись на благополучний результат і чудесну появу дружнього штучного інтелекту. І в цьому немає провини фахівців. Корінь проблеми не полягає в незнанні фахівців як створити дружню думаючу машину. Причиною, що може зумовити припинення буття людини, принаймні у звичних для нас формах, може стати переконання, що штучний інтелект буде обов'язково дружнім. Насамперед тому, що не можна нав'язати шляхи розвитку складній системі, нелінійному середовищу, які їй іманентно не властиві. Маючи справу зі створенням свідомості в машині, людина, як уже було сказано, не розуміє як працює вся система в цілому і тому їй відоме те віяло нелінійних векторів, атракторів, до яких буде прямувати штучний розум. Людині в принципі невідомо, що буде притаманне цій системі!

Віра у добрі наміри машинного інтелекту стає ще більш небезпечною після того, як машинний інтелект людського рівня, порушивши міру, через стрибок інтелектуального вибуху перейде у принципово нову якість – суперінтелект. Але все ж у своєму дослідженні Е. Юдковскі

висуває тезу про те, що дружність штучного інтелекту може базуватися на так званій функції корисності (синтез цінностей, уподобань, огорнутих у задоволення від досягнення мети, який впроваджений у визначення користі в алгоритмічних патернах [див.: 16].

Яку ж семантичну специфіку вкладають дослідники у поняття «дружній», вживаючи його в контексті штучного розуму? Передусім, штучний інтелект не повинен бути ніколи вороже чи амбівалентно налаштованим щодо людського виду, до якої мети б машина не прагнула і скільки б щаблів самовдосконалення не пройшла. Все це неможливо без глибокого розуміння машиною людської природи (чи розуміє її сама людина?), щоб надалі не заподіяти людині шкоди навіть у результаті випадкових, опосередкованих наслідків своїх дій (про що ми вже говорили в контексті осмислення Трьох законів А. Азімова). При цьому ми не хочемо одержати штучного розумного агента, який виконує короткострокові завдання за допомогою заходів, які б виявилися для людства шкідливими згодом.

Як приклад реалізації непередбачених наслідків, Н. Бостром пише про так звані згубні відмови [див.: 4, с. 153-176]. Тут ми наведемо деякі варіанти згубної відмови – порочну реалізацію й інфраструктурну надмірність. Перший зі сценаріїв, за яким штучний надрозум досягає поставленої мети у спосіб, оптимальний з його позицій, але суперечний загальнолюдській шкалі цінностей. Наприклад, бажаний результат, до якого хочу прийти я – постійна посмішка на моєму обличчі; спосіб досягнення цієї мети в розумінні машини – прямо задіяти нерв обличчя, що неминуче спричинить параліч міміки, в результаті чого посмішка не буде сходити з обличчя.

Найстрашніше, що вибір цього способу реалізації ніяк не зумовлений прагненням машини нашкодити людині. У цьому випадку порочна реалізація – маніпуляція на лицьовому нерві – для машинного інтелекту набагато оптимальніша, ніж звичні методи людини, в силу того, що це найповніший спосіб досягнення кінцевої мети. Чи існують якісь обхідні шляхи, котрі дозволяють розв'язати цю проблему? Можливо це вдасться завдяки конкретизації формулювання мети? Мета: без впливу прямо на лицьовий нерв забезпечити постійну посмішку. Спосіб побічної реалізації: активізація зон кори головного мозку, відповідальних за функції лицьового нерва. Вічно сяюча посмішка на обличчі забезпечена.

Можливо формулювання кінцевої мети утруднене внаслідок того, що ми використали звичний для людини понятєво-категорійний апарат? Спробуємо задати кінцеву мету, суть якої пов'язана з позитивним феноменологічним станом, щастям, суб'єктивним відчуттям комфорту, без опису поведінкових моделей. Гіпотетично ми припускаємо можливість реалізації фахівцями «обчислюваного» уявлення ідеї щастя та подальшого його вживлення в ембріон штучного інтелекту. Мета досить складна і спірна, якщо не неможлива. Тому в нашій статті шляхи її розв'язання досліджуватися не будуть. Але припустимо, що програмістам удалося поставити перед машинним розумом завдання ощасливити нас. Тоді порочна реалізація може набути такого вигляду: впровадження електродів у центри задоволення мозку.

Наступний приклад є формою реалізації іншого виду згубних відмов – інфраструктурна надмірність як такий процес, у якому розумний агент для досягнення конкретно поставленої мети перетворює всі відомі для нього види енергії, речовини й інформації в ресурс, виробничо-технічну базу для втілення цієї мети, внаслідок чого реалізація сутнісного потенціалу людства стає неможливою. В межах цього сюжету запрограмований штучний розум, якому як кінцеву мету було задано штампувати канцелярські скріпки, робить лише те, що від нього вимагалося, поза системою людських цінностей. У результаті суперінтелект перетворює доступний простір і речовину на фабрики з виробництва скріпок.

Відсутність закостенілих, догматичних цінностей – ще одна істотна, принципово важлива якість дружнього штучного інтелекту. Аксіологічні орієнтири машинної свідомості повинні зазнавати відповідних трансформацій у тісній кореляції зі змінами орієнтирів у суспільстві. Приміром, якби функція корисності гіпотетичного розумного агент була зорієнтована на ціннісні патерни переважної більшості населення Європи XVIII століття й не піддавалася змінам відповідно до розвитку суспільства, то й сьогодні цей штучний інтелект за одну з основ свого робочого алгоритму мав би систему архаїчних пережитків, серед яких рабовласництво, расова і статева дискримінація, публічні страти й інше. Система цінностей, впроваджених у дружню машину не повинна бути заданою раз і назавжди.

Такі теоретичні побудови виглядають, щонайменше, утопічно. З усього викладеного стає зрозуміло, що тема дружнього штучного інтелекту вимагає подальшої конкретизації та розвитку,

хоча її прихильники мислять украй оптимістично. Сьогодні наука не гарантує однозначної перспективи концепції дружнього надрозуму мовою математики, як і немає гарантії, що створення такого розуму взагалі можливе чи реальне його інтегрування в перспективні архітектури комп'ютерної свідомості. Але тепер, коли Трьом законам робототехніки заслужено надано статусу принципу побудови сюжету, а не засобу виживання, концепція дружнього машинного надрозуму, напевне, – це найкраще, що може запропонувати людство перед обличчям потенційної екзистенціальної загрози. Однак, дружня машина не створена, а проблем з її конструюванням цілком достатньо.

Одна з них пов'язана з тим, що велика кількість організацій в усьому світі працює над створенням штучного розуму рівня людини також у сфері суміжних технологій. У гонитві за пальмою першості жодна них не призупинить свою діяльність на цьому поприщі, чекаючи дня створення дружньої розумної машини. Занадто багато покладено на карту. Понад те, мало хто з учасників цієї гонитви залучений у науково-філософський дискурс, який стосується проблем дружнього штучного інтелекту.

До згаданих організацій, що працюють над створенням думаючої машини людського рівня, входять: AGIRI, CYC, Google, IBM (кілька проектів), LIDA, Nell, Numenta, SNERG і Vicarious. Також існує мінімум десять проектів, джерела фінансування яких не вважаються надійними: NARS, Novamente, Sentience, SOAR, а також DAPRA, яка підтримує прямо, або через посередників, проекти, пов'язані з розробкою штучного розумного агента, а також суміжні технології [див.: 13, с. 36]. Це далеко не повний перелік. У контексті цього виникає цілком логічний висновок: імовірність того, що перший штучний розум у межах легальних проектів побачить світ саме в лабораторіях MIRI мізерно мала, й отже, досить мізерна можливість впровадження в цю когнітивну архітектуру модуля дружності. Напевне, розроблювачів першої розумної машини буде мало хвилювати проблема дружності програмного забезпечення.

Однак існують і стратегічні вектори, які, можливо, надалі посприяють блокуванню ворожого надрозуму. Сьогодні існує освітня програма для передових університетів і математичних конкурсів, у межах якої MIRI та CEAR організували так звані «тренувальні табори розуму». В цих осередках навчають потенційних творців розумних програм і керівників, які формують подальшу технічну політику, інноваційному мисленню. У перспективі це повинно допомогти уникнути тупиків у розвитку та пасток на шляху вдосконалення створеного розуму. Зрозуміло, що цих заходів недостатньо, але MIRI та CEAR вдалося звернути увагу громадськості на важливий чинник ризиків, завдяки чому дедалі більше праць присвячується проблемі сингулярності, автори яких у такий спосіб розширюють і поглиблюють наукові та світоглядні аспекти феномену ризиків штучного розуму.

Та навіть якщо потенція стане реальністю і дружній розумний агент буде створений, немає жожної впевненості, що він залишиться таким після інтелектуального вибуху. Інакше кажучи, чи збереже штучний розум доброчесні якості, якщо його інтелект виросте в тисячі разів? Постійна поява нових якісних характеристик нескінченно зумовлюватиме новий характер кількісних змін. Результатом сотень, тисяч таких стрибків може стати трансформація осмислення феномену дружелюбності й однозначне спотворення запрограмованої семантики. У будь-якому випадку відбудеться зміна уявлень про мораль, а отже, спотвориться і функція корисності.

Е. Юдковскі з таким поглядом не згодний, вважаючи, що прогрес машинного інтелекту відіб'ється на поліпшенні ефективності функції корисності.

Реалізація такої можливості може мати місце в тому випадку, якщо процес інтелектуального вибуху пройде без флуктуацій, системного збою, природу яких ми в силу своєї антропоморфності навіть уявити собі не можемо. У людини і хробака великий відсоток спільного ДНК, але думка про те, що в нас можуть бути спільні цінності та мораль, як мінімум, смішна. Не змінила б цей стан і сенсаційна новина, в якій хробакові відводилася б роль нашого творця, який наділив своє дітище цінностями й ідеалами свого виду. Спочатку це спровокувало б у нас відторгнення і здивування, від якого ми швидко оправилися б і повернулися до свого звичного життя.

У цьому ж контексті дуже показовий приклад персонажу Доктора Манхеттена з графічного роману А. Мура, включеного в сотню кращих романів 20 століття [7]. Відомий фізик у результаті невдалого експерименту був розібраний на атоми і, відродившись, став іншою сутністю. Тепер йому під силу змінювати структуру матерії, природу простору-часу, він бачить минуле і майбутнє, його розум уже не турбує проблеми людства. Радості любові, страх смерті, суперечливість життя та інші «дріб'язки» – все те, без чого жодна людина не здатна себе помислити, речі, котрі є альфою і омегою

будь-якого особистісного існування, не турбують його більше. Після тривалих розмірковувань про сутність буття у всесвітньому масштабі він доходить висновку, що феномен життя занадто переоцінений. Надалі він розв'язує проблему загрози невідворотної ядерної війни, вбивши два мільярди чоловік, але тим самим, рятуючи сім мільярдів. Людський дружній інтелект розвинувся у Великий Розум, який зберіг якісь патерни людської моралі, але перетворився на щось чуже, позбавлене особистісних характеристик, по-своєму інтерпретуючи поняття оптимальності, допомоги, корисності. Іншими словами, у цьому випадку ми маємо справу з дружньою машиною, що пережила інтелектуальний вибух, у результаті якого розуміння «дружності» зазнало якісних трансформацій.

Д. Х'юз також висуває тезу про неспроможність ідеї стійких і незмінних первісних якостей штучного інтелекту в процесі стрибкоподібного розвитку машини, ґрунтуючись на осмисленні перетворення вітальних потреб людини (їжа, захист, самозбереження) на алгоритми, які відрізняються від первісних наборів цілей. Наприклад, людина може обрати стратегією свого життя аскетизм або целібат, що суперечить генетичній програмі нашого організму. Або середньостатистичний громадянин може стати терористом-смертником для того, щоб гонорар за теракт після смерті був оплачений родині. Тож людина здатна піддавати рефлексії свої цілі та вибудовувати різні алгоритми їхнього досягнення, які йдуть врозріз зі звичними моделями раціональності та базових інстинктів. У зв'язку із цим, думка про те, що створений штучний інтелект з відкритим і гнучким розумом (що, по суті, і є сутнісними характеристиками думаючої машини) не буде змінюватись у процесі розвитку, принаймні, наївна [див.: 3, с. 115].

Д. Хілліс також уважає, що людство поступово передає «кермо влади» комп'ютерам, тим самим віддаляючись від виробництва, управління, не вникаючи, по суті, у процес створення машинами ще більш складних машин та інших речей. Людина вже не розуміє, що і як відбувається. Технології створюють технології все більш самостійно без участі людини. На його думку, те що відбувається нагадує еволюцію найпростіших організмів у багатоклітинні. І в цій еволюції Д. Хілліс відводить людині роль амеби, яка не розуміє що і як ми створюємо [див.: 3, с. 117].

Істотною перешкодою на шляху до створення контрольованого штучного інтелекту стало недосконале програмування. Упевненість у непогрішності функціювання комп'ютерних програм розбивається об айсберг статистичних даних, що свідчать про те, що 60 млрд. доларів у рік американська економіка недоодержує в результаті дефектного програмування [див.: 13, с. 47]. Здивування викликає той факт, що комп'ютери як математичні машини повинні бути абсолютно лінійно детерміновані й передбачувані, тоді як у реальності все інакше: створення комп'ютерних програм, по суті, – одне з найімовірнісних інженерних завдань, полєднаних з багатьма помилками і проблемами безпеки.

С. Омохундро вважає, що способом боротьби проти неякісного програмування є створення систем, які усвідомлюють себе і здатні до рефлексії над своєю поведінкою в процесі досягнення поставлених цілей. Інакше кажучи, вони повинні вміти саморозвиваються. Програми саморозвитку – неодмінна якість і неминучий етап на шляху до створення думаючої машини. Однак програмне забезпечення, що усвідомлює себе, сьогодні ще не розроблене, а програми, що модифікують себе, досить поширені [див.: 9].

Одним з алгоритмів машинного навчання, який використовує можливості природного добору для пошуків відповідей, є генетичне програмування. Цей алгоритм – найважливіший інструментарій для написання потужних програм. У цьому виді програмування, на відміну від звичайного програмування, де використовується людська логіка, застосовується логіка комп'ютерна. У звичайному програмуванні при написанні рядка коду використовується робота програміста, завдяки чому процес обробки даних у межах схеми «вхід-вихід» прозорий і підвладний верифікації. У випадку застосування генетичного програмування програміст лише описує завдання, а її розв'язання передається на волю природного добору. Генетична програма генерує фрагменти коду, які є елементами системи наступного покоління. Найпрогресивніші з них синтезуються випадково, створюючи нову генерацію. Програма тим більш оптимальна і перспективна, чим більше вона зуміла наблизитися до розв'язання поставленого перед нею завданням. Такий процес у комп'ютерній еволюції, по суті, є подобою природного добору в природі, упродовж якого слабкі відкидаються, а кращі вступають у взаємодію знову, приводячи до випадкових трансформацій окремих команд і змінних. Такі якісні стрибки у розвитку природи називаються мутаціями. Програміст, запустивши таку генетичну програму, надалі від корекції в її роботі може усунутися. По суті, якийсь компьютинговий деїзм, у якому програмістові відводиться

роль ініціатора, який запустив процес, але надалі не впливає на причини, наслідки яких у перспективі, можливо, запустять ланцюг подій, здатних породити нову реальність.

С. Омохундро, ілюструючи небезпеку та непередбачуваність роботи систем, здатних змінювати себе, наводить приклад робота-шахіста. Звичайно ж, мова не йде про комп'ютерну програму, встановлену в будь-якому комп'ютері середньої продуктивності. С. Омохундро має на увазі потенційного робота-шахіста, здатного до самоусвідомлення та самовдосконалення. Як поводитиметься цей розумний агент у ситуації, якщо ви зіграєте з ним партію в шахи, а потім дасте йому команду відключитись? І вся проблема полягає в тому, що відключення для машини, яка усвідомлює себе, – подія досить значима в силу того, що включитися самостійно вона не може. Тому вкрай важлива впевненість у цілком адекватній оцінці реального стану справ. І в прагненні оцінити ситуацію якнайглибше робот може прийти до думки виділити якісь ресурси на пізнання природи реальності перед тим, як вийти з неї, відключившись.

Виникає дуже важливе питання: яка кількість ресурсів розумний агент може порахувати достатнім? Відповідь, що дав С. Омохундро змусить задуматися багатьох оптимістів створення думаючих роботів: «Роботизированная програма вирішити рішення на це все доступні людині ресурси» [9].

Отже, нелінійний простір розвитку штучного інтелекту – відображення нелінійної еволюції людини, яка у боротьбі за ресурси не раз ставила себе на грань виживання, але вперше за десятки тисяч років свого існування людині доведеться мати справу із системами, які переважаючими її в іграх, рівних у яких їй досі не було, а саме в іграх розуму. Розкоші у вигляді десятків-сотні років для техно-гуманітарної адаптації може не виявитися, в силу того, що самовдосконалюваний штучний розум вийде у своєму розвитку за межі здатності людини якось впливати на хід подій. А штучний інтелект, легко подолавши безодню між штучним і природним, посяде почесне місце в природній ніші, впливаючи на процеси Всесвіту, або просто замінить його. Доля людини в цьому контексті досить невизначена. С. Шостак справедливо зазначив, що цивілізації самі створюють собі спадкоємців [див.: 15].

Очевидно, в Людини є два варіанти: думати, що ще є час інтелектуально й стратегічно «згрупуватися», зробивши все можливе, щоб приборкати «чорно-білого мустанга» за назвою Штучний Інтелект, що стрімко мчиться за горизонт подій сингулярності. Другий – сподіватися на те, що С. Шостак недостатньо прозорливий.

Література

1. Азимов А. Загадки мироздания. Известные и неизвестные факты / Айзек Азимов. – М.: Центрполиграф, 2016. – 432 с.
2. Азимов А. Я - Робот / Айзек Азимов. – М.: Эксмо, 2007. – 1296 с.
3. Баррат Д. Последнее изобретение человечества: Искусственный интеллект и конец эры Homo sapiens / Джеймс Баррат. – М.: Альпина нон-фикшн, 2015. – 304 с.
4. Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии / Ник Бостром. – М.: Манн, Иванов и Фербер, 2015. – 496 с.
5. Курцвейл Р. Эволюция разума / Рэй Курцвейл. – М.: Эксмо, 2016. – 448 с.
6. Ллойд С. Программируя Вселенную: Квантовый компьютер и будущее науки / Сет Ллойд. – М.: Альпина нон-фикшн, 2014. – 256 с.
7. Мур А. Хранители / Аллан Мур. – М.: Азбука, Азбука-Аттикус, 2014. – 528 с.
8. Назаретян А. П. Нелинейной будущее / А.П. Назаретян. – М.: МБА, 2013. – 440 с.
9. Омохундро С. Чем нам угрожают роботы? [Електронний ресурс] // Technowars. – 2015. – Режим доступу до ресурсу: https://technowars.defence.ru/article/1468/.
10. Стибел Д. Почему я нанимаю неудачников [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: http://www.michelino.ru/2016/06/blog-post_11.html.
11. Тест Тьюринга пройден (на детском уровне) [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: https://habrahabr.ru/post/225599/.
12. Тьюринг А. Может ли машина мыслить? / Алан Тьюринг. – СПб.: Эдиториал УРСС, Ленанд, 2016. – 128 с.
13. Форд М. Технологии, которые изменят мир / Мартин Форд. – М.: «Манн, Иванов и Фербер», 2014. – 268 с.
14. Шелли М. Франкенштейн, или Современный Прометей / Мэри Шелли. – СПб.: Мартин, 2015. – 256 с.
15. Шостак С. Вероятно внеземной разум существует. Вы готовы? [Електронний ресурс]. – 2013. – Режим доступу до ресурсу: http://www.fassen.net/video/gLJZJuPM24M/.
16. Юдковски Е. Обратное глупости не есть ум [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: http://lesswrong.ru/w/Обратное_глупости_не_есть_ум.

Стаття надійшла до редакції: 22.11.2016

References

1. Azimov A. (2016), *Zagadki mirozdaniâ. Izvestnye i neizvestnye fakty*, Centrpoligraf, M., 432 s.
2. Azimov A. (2007), *Â – Robot*, Èksmo, M., 1296 s.
3. Barrat D. (2015), *Poslednee izobretenie čelovečestva: Iskusstvennyj intellekt i konec èry Homo sapiens*, Al'pina non-fikšn, M.:, 304 s.
4. Bostrom N. (2015), *Iskusstvennyj intellekt. Ètapy. Ugrozy. Strategii*, Mann, Ivanov i Ferber, M., 496 s.
5. Kurcvejl R. (2016), *Èvolûciâ razuma*, Èksmo, M., 448 s.
6. Llojd S. (2014), *Programmiruâ Vselennuû: Kvantovyj komp'ûter i budu*ŝ*ee nauki*, Al'pina non-fikšn, M., 256 s.
7. Mur A. (2014), *Hraniteli*, Azbuka, Azbuka-Attikus, M., 528 s.
8. Nazaretân A. P. (2013), *Nelinejnoj budu*ŝ*ee*, MBA, M., 440 s.
9. Omohundro S. (2015), *Čem nam ugrožaût roboty?*, available at: https://technowars.defence.ru/article/1468/ (accessed August 31, 2016)
10. Stibel D. (2015), *Počemu â nanimaû neudačnikov*, available at: http://www.michelino.ru/2016/06/blog-post_11.html (accessed August 31, 2016)
11. *Test T'ûringa projden (na detskom urovne)*, available at: https://habrahabr.ru/post/225599/ (accessed August 31, 2016)
12. T'ûring A. (2016), *Možet li mašina myslit'?*, Èditorial URSS, Lenand, SPb.,128 s.
13. Ford M. (2014), *Tehnologii, kotorye izmenât mir*, «Mann, Ivanov i Ferber», M., 268 s.
14. Šelli M. (2015), *Frankenštejn, ili Sovremennyj Prometej*, Martin, SPb., 256 s.
15. Šostak S. (2013), *Veroâtno vnezemnoj razum su*ŝ*estvuet. Vy gotovy?*, available at: http://www.fassen.net/video/gLJZJuPM24M/ (accessed August 31, 2016)
16. Ûdkovski E. (2015), *Obratnoe gluposti ne est' um* available at: http://lesswrong.ru/w/Обратное_глупости_не_есть_ум (accessed August 31, 2016)

Received: 22.11.2016

Summary

Snehiriov I. The Nonlinear Space of Artificial Intelligence: the Philosophical and Ideological Understanding. *The author from the perspective of a nonlinear prediction methodology reveals the prospects and risks of artificial intelligence. The reasons as the optimistic forecasts and complex threats posed by "horizons" singularity, due to the creation of machine intelligence. Particular attention is paid to the concept of a friendly intelligent machine in terms of intellectual explosion. One of the main factors influencing the relatively stable development of man and social systems is transformational activity, which, as the forms, means and methods of innovation become more complex, gradually turned into technology. Their negentropic potential and technical implementation in various spheres of life since ancient times had a significant impact on the formation of scientific rationality of the modern type, which undoubtedly correlates with the emergence of the information society with its characteristic increase in the role of statistical regularities and the nonlinearity factor. From the middle of the 20th century, science began to play a leading role in the system of social production, and high technology began to claim the role of a new attractor, determining new epistemological vectors and axiological horizons for the development of social systems. But the more complex a social system, the more it is subject to the influence of stochastic factors that create a nonlinear space for further development paths that are formed as a result of the actualization of phase transitions-the forced responses of a nonequilibrium structure to the threat of a decrease in sustainability.*

In this context, progress is not an end in itself, not of self-imposed value, but acts as a way of preserving a relatively complex integrity. The ways of achieving such a state are nonlinear, since in advance to calculate their number, the degree of determination and danger for any period of time is not possible. The future can become like the "better" of the present in strictly defined parameters, and "worse" in other parameters and technologies, especially science-intensive ones, which can cause an existential crisis of global proportions play a significant role in this sense. Solving some contradictions "launches" a nonlinear chain of many other, new, even more ambiguous problems. In the future this causes the emergence of vectors of evolutionary changes: from more stochastic ("natural") to less probable states. In accordance with the nonlinear model, progress as "removal from the natural niche" means the restoration of the relative stability of the system at an increasingly higher level of disequilibrium. An attempt to philosophical reflection several options disastrous failure, namely the realization of evil and infrastructure redundancy. <u>Keywords</u>*: artificial intelligence, the singularity, nonlinearity, anthropomorphism, "God in a box," Turing test, the Three Laws of Robotics, "black box", genetic algorithms.*