Borys Grinchenko Kyiv Metropolitan University
Faculty of Romance and Germanic Philology
Linguistics and Translation Department

Translation project:
***Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI* by Karen Hao**

Перекладацький проєкт:
Переклад книги К. Гао «Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI»

BA Paper

Karina Dovbysh
PERb12240d

Цим підписом засвідчую,
що подані на захист рукопис та
електронний документ є ідентичні

27.05.2026 р.

Research supervisor:
Professor O. Komar

Kyiv 2026

Abstract

This translation project focuses on an excerpt from the book *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI* by Karen Hao. The aim is to identify the challenges involved in translating artificial intelligence terminology from English into Ukrainian. Our study found that the source text is technical and analytical, impacting translation decisions. Its main characteristics include specialised terminology, neologisms and context-dependent technical meanings of common lexical units. These features present translation challenges, including the absence of standardised equivalents in Ukrainian and semantic complexity. To address these challenges, the terminology in the text was classified according to the system proposed by A. Sydor and R. Nanivsky, and the translation techniques were analysed using the classification of L. Molina and A. Albir. The analysis showed that the most common techniques are established equivalents, borrowing, calque, modulation and description.

Keywords: *AI discourse, artificial intelligence, non-fiction, technical translation, translation techniques.*

Анотація

Цей перекладацький проєкт присвячений уривку з книги *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI* авторства Карен Гао. Мета проєкту полягає у визначенні проблем, пов'язаних з перекладом термінології штучного інтелекту з англійської на українську. Дослідження показало, що текст оригіналу має технічний та аналітичний характер, що впливає на перекладацькі рішення. До його основних особливостей належать спеціалізована термінологія, неологізми та залежні від контексту технічні значення загальновживаних лексичних одиниць. Ці особливості створюють перекладацькі труднощі, зокрема відсутність усталених еквівалентів в українській мові та семантичну складність. Для вирішення цих проблем термінологію, що міститься в тексті, було класифіковано за запропонованою А. Сидором та Р. Нанівським системою, а також проаналізовано перекладацькі прийоми на основі класифікації Л. Моліни та А. Альбір. Аналіз показав, що найпоширенішими прийомами є усталені еквіваленти, запозичення, калька, модуляція та описовий переклад.

Ключові слова: *наукова література, перекладацькі прийоми, технічний переклад, ШІ-дискурс, штучний інтелект.*

Contents

Introduction

In modern society, the accelerated development of technology has a significant impact on all areas of human activity, particularly linguistics. The fields of information technology and artificial intelligence have fundamentally transformed the terminology used to describe and interact with them. Discourse in the field of artificial intelligence is characterized by the constant emergence of new concepts, mixed terminology and context-dependent meanings, which poses major challenges for the translation process. These challenges are particularly evident in the translation of contemporary analytical and narrative texts on the development of companies such as OpenAI and the broader field of artificial intelligence, where technical, social, and ideological aspects are closely intertwined.

One of the main challenges in the discourse on artificial intelligence is the ambiguity of certain terms and the contextual nature of widely used lexical units. These aspects require a careful approach during translation, as an unclear rendering of technical terminology can lead to a loss of meaning and a misinterpretation of a concept's function. Additionally, the dynamic nature of this field leads to the constant emergence of neologisms and the absence of standardized Ukrainian equivalents for many terms.

A number of scholars have studied the problem of translating technical and artificial intelligence terminology. Notable researchers such as P. Newmark and J. Byrne examined issues in technical translation and terminology, emphasising the importance of accuracy and functional adequacy in translating specialised texts. Ukrainian scholars, such as A. Sydor and R. Nanivsky, have proposed classifications of terminology for more detailed analysis of vocabulary in specialised discourse. These approaches became the foundation of our research.

The object of the research is artificial intelligence discourse in non-fiction texts.

The subject is the artificial intelligence terminology and peculiarities of its translation from English into Ukrainian.

The aim of our translation project is to analyse the specific features of translating technical terminology in the field of artificial intelligence, using non-fiction text as an example, and to identify effective strategies for translating these language units into Ukrainian.

The aim of our study involves the following **objectives:**

- to analyse the types and structural characteristics of artificial intelligence-related terms in the selected fragment.
- to identify the main difficulties arising in the process of translating artificial intelligence technical terminology.
- to define translation strategies and techniques used for rendering AI terms into Ukrainian.
- to conduct a practical analysis of translation examples from the selected text and determine their accuracy.

The structure and the body of the research. The translation project consists of an introduction, two chapters, conclusions, references, and appendices. The source text contains 44,456 characters, while the target text contains 46,935 characters. The total number of pages is 53.

Chapter 1. Translation of Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI by Karen Hao

Source Text

Chapter 10

Gods and Demons

To live in San Francisco and work in tech is to confront daily the cognitive dissonance between the future and the present, between narrative and reality.

The first time I moved to San Francisco, as a university sophomore for a summer internship, I was dazzled by the quaint aesthetics of the city. The colorful Spanish-style architecture, the limited number of skyscrapers, the hills steep enough to make driving a stick shift a test of reflexes. There was an endless supply of perfectly ripe avocados and toasted sourdough bread and smooth Blue Bottle lattes. There were different neighborhoods, all with their own look and culture.

When I returned full time after graduation to work at a tech startup, I crammed into a three-bedroom apartment with three other roommates in the Castro. On weekends we would hike across the

Target Text

Розділ 10

Боги та демони

Життя у Сан-Франциско й робота у сфері технологій надає можливості щодня стикатися з когнітивним дисонансом між майбутнім і сьогоденням, між ідеєю та реальністю.

Вперше я переїхала до Сан-Франциско на другому курсі — на літнє стажування. Мене зачарувала особлива краса міста: барвиста архітектура в іспанському стилі, всього декілька хмарочосів та круті схили, на яких керувати авто з «механікою» було справжнім випробуванням. Стиглі авокадо, підсмажений хліб на заквасці та ніжне лате з Blue Bottle тут завжди були у доступі. Кожен район здавався окремим світом.

Після випуску я повернулася працювати в техстартап уже на повну зайнятість і оселилася в трикімнатній квартирі в районі Кастро разом із трьома сусідками. У вихідні ми гуляли пагорбами й

rolling hills and forage from public fruit trees. On weeknights, neighbors—all young twentysomethings in the tech industry—would pop over unannounced to play board games, drink wine, and while away the evenings. House parties were a constant, as were weekend trips to stunning nature: Lake Tahoe in the north, Big Sur to the south, tall, majestic redwoods everywhere around us. Life was easy. We were young, making salaries relatively standard in the tech industry that placed us nationally in our age group's top 5 percent.

But there was that dissonance. On the way to work, I would pass people shooting up drugs in front of the subway stations, the unhoused peeing on sidewalks just blocks from my office. Meanwhile, our startup's chef, playfully named "the happiness engineer," would cook or cater an abundance of food for our free office lunches. Leftovers often went straight into the trash. If we stayed late, we got free dinner—and were emphatically implored to take an Uber home for safety reasons. It was all too easy for the privileged to grow accustomed to moving through the city in ways that shielded them from seeing the realities of how the other half lived.

The dichotomy encapsulated how the tech industry could profess big, bold visions about changing the world and building a

зривали плоди з фруктових дерев. Будніми вечорами молоді сусіди, що також працювали у сфері технологій, часто забігали без попередження, щоб пограти в настільні ігри, випити вина й просто «вбити» трохи часу. Домашні вечірки відбували постійно, як і наші поїздки на природу: до озера Тахо на півночі, у Біг-Сюр, де навколо росли високі величні секвої, на півдні. Ми були молодими й мали типову для техсфери зарплатню, завдяки якій перебували у топ 5 найбагатших людей нашого віку в країні. Життя було простим і безтурботним.

Та в цьому всьому було щось суперечливе. Дорогою на роботу я бачила людей, які вживали наркотики просто біля входу в метро, безхатьків, що мочилися на тротуарах за кілька кварталів від мого офісу. Тим часом у нашому стартапі шеф-кухар, якого жартома називали «інженером щастя», готував величезну кількість їжі для безкоштовних офісних обідів, залишки якої часто опинялися у смітнику. Коли ми працювали допізна, нам пропонували безкоштовну вечерю і, з міркувань безпеки, наполегливо просили їхати додому на таксі. Людям із привілеями було надзвичайно легко звикнути до способу життя, який приховував від них існування іншої частини міста.

Ця дихотомія яскраво демонструвала, як техіндустрія здатна декларувати масштабні ідеї про зміну світу та створення кращого

better future while ignoring the very problems at its door. It was a dichotomy that Altman would sometimes comment on in his own way—getting right up to yet never fully acknowledging the utter contradiction of declaring the problem of creating and managing beneficial AGI possible, but San Francisco's housing crisis too tough to tackle. " Where I grew up, no one would ever walk by a person collapsed on the side of the street on their way to work and not do something about it," he once said, comparing suburban St. Louis to San Francisco. "I do blame the tech industry for a lot of things that have gone wrong with the city, but not all of them. But we have, just over time, had this, like, unbelievable wealth generation in this small geographic space, in this small period of time, and I think not been particularly thoughtful about the effects of that on the community as a whole. And because those problems are so hard and so hard to think about, I think most people just choose not to, and they just accept this."

майбутнього, водночас не помічаючи проблем поряд. Про це іноді говорив і Альтман — по-своєму підходячи до питання, але так і не визнаючи повної суперечності між вірою в можливість створення й керування корисним загальним штучним інтелектом (AGI) та твердженням, що житлову кризу у Сан-Франциско вирішити нібито надто складно. «Там, де я виріс, ніхто б не пройшов повз людину, що дорогою на роботу знепритомніла на узбіччі, не зробивши бодай чогось», — якось сказав він, порівнюючи передмістя Сент-Луїса із Сан-Франциско. «Я достоту звинувачую техіндустрію в багатьох проблемах міста, але не в усіх. За відносно короткий час у відносно невеликому географічному просторі відбулося неймовірне накопичення багатства, й, як мені здається, ми не дуже замислювалися над тим, як це впливає на спільноту загалом. А оскільки ці проблеми такі складні для осмислення, більшість людей просто не думає про них і приймає це як буденність».

It was in this context that effective altruism arrived from the UK and found its most loyal audience. EA, to which many in OpenAI's Safety clan were early adherents, made for the perfect Silicon Valley ideology. It preaches making the world a better place and doing it with rigorous logic, being disciplined enough to focus on the far future instead of the present, and fervently embracing the

У цьому середовищі ефективний альтруїзм (EA), що прийшов із Великої Британії, знайшов свою найвідданішу публіку. Багато представників безпекового напряму OpenAI долучилися до руху ще на ранньому етапі, і EA швидко перетворився на зручну ідеологію для технологічного середовища. Він проповідує прагнення зробити світ кращим — але з опорою на сувору логіку;

principles of capitalism and libertarianism—all in the name of morality.

Core to the EA philosophy is a mathematical concept called "expected value." The expected value of something is calculated by multiplying the probability that it will occur with its quantified positive or negative impact. It's a tool that can lead to counterintuitive thinking. In a 2013 paper, EA cofounder William MacAskill, at the time a doctoral student who would become an Oxford philosophy professor, argued, based on this logic, that it was more altruistic in the long run to take a more morally ambiguous job to get rich and donate that money through optimized philanthropy than to commit to a life of working for a morally good charity. Based on his conservative estimates, he wrote, the expected value of being a rich philanthropist would in fact be forty times greater than being an ascetic charity worker. He laid out the math based on a series of arbitrary numbers: graduates who worked to get rich might on average fund two charity workers, each working at charities ten times more cost-effective than one they would have otherwise worked for. Half of the benefits they produced if they chose the charity route would also happen with or without them anyway. His argument

закликає до дисципліни мислення й зосередженості на далекому майбутньому замість сьогодення і водночас активно приймає капіталізм і лібертаріанські цінності — й усе це під прапором моралі.

Ключовим елементом філософії EA є математична концепція «очікуваної цінності». Її обчислюють, множачи ймовірність події на кількісно оцінений позитивний або негативний результат. Це інструмент, що здатен привести до доволі парадоксальних висновків. У статті 2013 року співзасновник руху Вільям МакАскілл, тодішній аспірант, який згодом став професором філософії в Оксфорді, стверджував, що з погляду довгострокового альтруїзму доцільніше обрати морально неоднозначну кар'єру, розбагатіти й скеровувати ці кошти на стратегічно оптимізовану філантропію, ніж присвятити життя роботі в етично бездоганній благодійній організації. За його консервативними підрахунками, очікувана цінність заможного філантропа була б у сорок разів вищою, ніж у скромного працівника благодійності. Свої розрахунки він будував на низці умовних припущень: випускники, які прагнули заробити великі статки, могли б у середньому фінансувати двох працівників благодійних організацій, кожен із яких працював би в структурах удесятеро ефективніших за ті, де вони могли б працювати самі. Крім того, половина користі від їхньої благодійної

would be encapsulated in one of the movement's most popular mantras: "Earn to give."

Under the logic of expected values, the founding EA philosophers also developed a framework for identifying the highest priority problems. Such problems need to be "big in scale," boosting their expected value; "tractable," possible to fix for proportionally little time or money; and "unfairly neglected," suffering severe and disproportionate underinvestment. While the movement encourages people to use the framework to identify their own problems, it also has recommendations of which problems it deems most worthy. " I and others in the effective altruism community have converged on three moral issues that we believe are unusually important, score unusually well in this framework," MacAskill said in a TED Talk in 2018.

First is improving global health, such as by distributing cheap yet effective bed nets to prevent malaria. Second is abolishing factory farming, which could improve billions of animals' lives "for just pennies per animal." Third is existential risks: risks that have a

діяльності однаково виникла б незалежно від їхньої участі. Цей аргумент зрештою зводився до одного з найвідоміших гасел руху — «Заробляй, щоб віддавати».

У межах концепції очікуваної цінності філософи-засновники ЕА сформували також критерії визначення пріоритетних проблем. Ці проблеми мають бути «масштабними» — адже це підвищує їхню очікувану цінність; «розв'язуваними» — тобто такими, які можна вирішити за рахунок відносно невеликих витрат часу чи коштів; і «несправедливо проігнорованими» — такими, що отримують непропорційно мало уваги та фінансування. Хоча рух заохочує кожного застосовувати ці критерії для самостійного визначення важливих напрямів, він водночас пропонує власний перелік проблем, які вважає найпріоритетнішими. «Я та інші представники спільноти ефективного альтруїзму дійшли згоди щодо трьох моральних питань, які, на нашу думку, є надзвичайно важливими й особливо добре відповідають цим критеріям», — сказав МакАскілл під час виступу на TED у 2018 році.

По-перше – покращення глобального здоров'я — наприклад, через поширення дешевих, але ефективних протимоскітних сіток для запобігання малярії. По-друге — скасування індустріального тваринництва, що могло б покращити життя мільярдів тварин

dramatically high expected negative value because—no matter how improbable—they could destroy all of humanity and cut short all of the future value that would otherwise be generated for the rest of civilization. In this third category are further recommendations for what constitutes an existential risk: global pandemics, nuclear war, and rogue artificial intelligence. With the identification of theoretical rogue AI as an existential risk, EA promulgated the same brand of AI safety that had been entwined within OpenAI's DNA from the very beginning and had played a critical role in The Divorce. Amodei and his fellow Anthropic cofounders fundamentally disagreed with Altman and the other OpenAI executives over how seriously to take the possibility of AI devastating civilization. Amodei, who took it very seriously, viewed Altman's behaviors—his lack of transparency on the Microsoft deal; his apparent compulsion to always tell people what they wanted to hear to gain their agreement, only for them to discover the misdirection too late—not just as the typical machinations of a Silicon Valley executive but as alarming, immoral behavior that could jeopardize the fate of humanity. As Anthropic established itself, it would lean into this reputational distinction: Where Altman's OpenAI was toying recklessly with humanity's future, Anthropic was the principled, AI-safety-first company.

In 2021, as the Amodei siblings announced Anthropic,

«усього за копійки». По-третє — екзистенційні ризики: загрози з надзвичайно високою очікуваною негативною цінністю, адже, якими б малоймовірними вони не були, їхні наслідки здатні знищити людство й обірвати розвиток цивілізації. До таких ризиків зараховують глобальні пандемії, ядерну війну та неконтрольований штучний інтелект. Визнавши гіпотетичний вихід ШІ з-під контролю екзистенційною загрозою, ЕА почав просувати той самий підхід до безпеки ШІ, що був закладений в OpenAI від самого початку і відіграв ключову роль у подіях, відомих як «Розлучення». Амодей та інші співзасновники Anthropic принципово розходилися з Альтманом і керівництвом OpenAI в оцінці того, наскільки серйозною може бути загроза, яку ШІ становить для цивілізації. Амодей, який сприймав її вкрай серйозно, вбачав у поведінці Альтмана — непрозорості щодо угоди з Microsoft та схильності говорити людям те, що вони хотіли почути, аби заручитися їхньою підтримкою, — не просто типові маневри керівника з Кремнієвої долини, а тривожну й аморальну позицію, здатну поставити під загрозу долю людства. Вибудовуючи власну репутацію, Anthropic зробила цю відмінність принциповою: якщо OpenAI Альтмана безрозсудно бавилося з майбутнім людства, то Anthropic прагнула постати компанією принципів і пріоритету безпеки ШІ.

У 2021 році, коли брат і сестра Амодеї оголосили про

interest in this catastrophic and existential AI safety ideology was accelerating, chiefly due to EA's rapidly expanding sphere of influence. EA had grown from a niche philosophy into a mainstream movement through an influx of cash from tech billionaires.

A decade earlier, Facebook cofounder Dustin Moskovitz and his wife, former journalist Cari Tuna, had formed a nonprofit called Good Ventures to give away most of their fortune. At the time, Holden Karnofsky, Daniela Amodei's future husband, had been running a different organization called GiveWell, which he'd founded in 2007 after leaving the hedge fund Bridgewater Associates. With a shared desire to distribute money with evidence-based methods, Good Ventures and GiveWell formed a partnership in 2011, which they later named Open Philanthropy. They began ramping up funding to the key issue areas that MacAskill had recommended—its grants toward AI safety research in particular were guided by the EA framework. Open Philanthropy became an independent organization in June 2017.

More recently, a new tech billionaire had entered the scene: Samuel Bankman-Fried, a rapidly rising star for his wild success cofounding the crypto exchange FTX and crypto trading firm

створення Anthropic, інтерес до екзистенційно-катастрофічної парадигми безпеки ШІ стрімко зростав — передусім завдяки швидкому розширенню впливу ефективного альтруїзму. ЕА, що починався як нішова філософія, перетворився на мейнстримний рух завдяки потужному припливу коштів від технологічних мільярдерів.

За десять років до цього співзасновник Facebook Дастін Московіц разом із дружиною — колишньою журналісткою Карі Туною — створили благодійний фонд Good Ventures, щоб передати на доброчинність більшу частину свого статку. Тоді ж Голден Карнофскі — майбутній чоловік Даніели Амодей — керував іншою організацією, GiveWell, яку заснував у 2007 році, залишивши хедж-фонд Bridgewater Associates. Прагнучи розподіляти кошти на основі доказових методів, Good Ventures і GiveWell у 2011 році уклали партнерство, яке згодом отримало назву Open Philanthropy. Вони почали активно спрямовувати фінансування на ключові напрями, рекомендовані МакАскіллом, зокрема на дослідження з безпеки ШІ, що здійснювалися відповідно до рамки ефективного альтруїзму. У червні 2017 року Open Philanthropy стала незалежною організацією.

Пізніше до них долучився новий техномільярдер — Семюел Бенкман-Фрід, який здобув славу завдяки шаленому успіху криптобіржі FTX та трейдингової компанії Alameda Research,

Alameda Research. Bankman-Fried, or SBF as he is known, credited EA for his origin story. A physics major at MIT, he said he had wanted to be an academic before MacAskill convinced him over lunch of the moral superiority of “earn to give.” SBF subsequently set his course on making himself as rich as possible in order to eventually, he pledged, put it all into philanthropy.

As he amassed his wealth in remarkably short order, SBF donated tens of millions to political candidates, both Democrat and Republican, including the first ever EA-backed candidate in 2022 in Oregon’s Sixth Congressional District (who ultimately didn’t win the primary). SBF’s exchange inked lavish deals totaling billions on sports marketing involving top athletes like Tom Brady and Steph Curry and top sports like Formula One. Into the EA movement, he pumped not just money but star power. The richer and more famous he became, the more he raised the profile of the ideology and its cofounder MacAskill. At the start of 2022, SBF announced the creation of his own EA-driven philanthropic project, FTX Future Fund, to distribute at least $100 million and up to $1 billion by the end of the year.

співзасновником яких він був. Бенкман-Фрід, відомий як СБФ (Сем Банкман-Фрід) називав ефективний альтруїзм частиною своєї історії становлення. Він згадував, що в якості студента-фізика MTI (Массачусетський технологічний інститут) спершу планував академічну кар’єру, аж поки під час обіду МакАскілл не переконав його в моральній перевазі принципу «заробляй, щоб віддавати». Після цього СБФ узяв курс на те, щоб стати якомога багатшим і, як він обіцяв, згодом спрямувати весь статок на філантропію.

Накопичивши багатство вражаючими темпами, СБФ почав жертвувати десятки мільйонів доларів політичним кандидатам — як демократам, так і республіканцям, зокрема першому в історії кандидатові, відкрито підтриманому рухом ЕА, на виборах 2022 року в Шостому окрузі штату Орегон (який зрештою програв праймеріз). Його біржа уклала багатомільярдні контракти у сфері спортивного маркетингу — із зірками масштабу Тома Брейді та Стефа Каррі, а також із такими брендами, як «Формула-1». У рух ефективного альтруїзму СБФ вкладав не лише гроші, а й статус. Чим багатшим і відомішим він ставав, тим більше підвищував авторитет ідеології та її співзасновника Макаскілла. На початку 2022 року SBF оголосив про запуск власного філантропічного проєкту, заснованого на принципах ЕА, — FTX Future Fund, який планував розподілити щонайменше 100 мільйонів і ще близько 1 мільярда

In large part due to Open Phil and FTX Future Fund, 2021 and 2022 saw a jump in cash flow to EA-backed AI safety research. According to estimates compiled by a member of the EA community and Open Phil data, funding leapt up above $100 million each for both years, after averaging less than half that amount over the previous seven years. The influx fueled and was fueled by a proliferating belief that the dramatic leap in capabilities from GPT-2 to GPT-3 made preventing theoretical rogue AI and existential AI risks more urgent than ever before. More and more people flocked to these kinds of AI safety projects, drawn in by the financial incentive or by ideology, as membership in the broader EA movement ballooned. EA had long touted the importance of pandemic preparedness, and now, in the midst of an actual pandemic, its remarkable prescience won it new adherents. The psychological toll of a global catastrophe had also left many people anxious and unmoored, searching for purpose.

The growing membership in the AI safety community, which

доларів до кінця року.

Переважно завдяки Open Phil та FTX Future Fund у 2021–2022 роках різко зросло фінансування досліджень безпеки ШІ, спонсорованих ЕА. За підрахунками, здійсненими одним із представників спільноти ЕА на основі даних Open Phil, обсяг фінансування в кожному з цих років перевищив 100 мільйонів доларів, тоді як упродовж попередніх семи років він у середньому становив менш ніж половину цієї суми. Цей приплив коштів супроводжувався — і водночас посилював — переконання, що драматичний стрибок у можливостях від *GPT-2* до *GPT-3* зробив запобігання теоретичному виходу ШІ з-під контролю та екзистенційним ризикам нагальнішим, ніж будь-коли. До проєктів із безпеки ШІ долучалося дедалі більше людей — когось приваблювали фінансові можливості, когось ідеологічні переконання, — а чисельність ширшого руху ЕА стрімко зростала. Ефективний альтруїзм завжди наголошував на важливості готовності до пандемій, і тепер, у розпал реальної глобальної кризи, ця майже пророча передбачливість принесла рухові нових прихильників. Психологічний тягар глобальної катастрофи також залишив багатьох людей у стані тривоги й розгубленості — у пошуках сенсу та опори.

Зростання спільноти з безпеки ШІ, яка поєднала підходи

knit together EA-backed AI safety with other strains of catastrophic, existential, and risk-focused thinking, swelled Anthropic's ranks just as it restocked OpenAI's Safety clan. Online EA and AI safety forums, the primary ground for the overlapping movements to propagate, exchange, and debate ideas, encouraged adherents to work at the major AI labs, especially those they felt needed more AI safety watchdogs, like OpenAI and DeepMind, to shape and mold their trajectory. The influx of members in AI safety also popularized the community's lexicon more broadly in the AI industry. How fast you think AI will advance and reach major milestones like AGI is your "AI timeline." How likely you think it is that AGI will lead to catastrophic outcomes, meaning the killing off of *most* of the human population, or existential outcomes, meaning the complete and total extinction of humanity, is your "p(doom)," short for *probability of doom*. "Hardware overhang," as referenced in OpenAI's 2021 research road map, is another dictionary entry, as is "AI takeoff," the process of AGI improving to the point of superintelligence and thus capable enough to outwit humanity. "Acceleration risk" refers to the risk of triggering a heightened competition between companies or countries that leads to a potentially dangerous acceleration of AI advancement and a shortened AI timeline.

руху ефективного альтруїзму з іншими течіями катастрофічного, екзистенційного та ризик-орієнтованого мислення, водночас розширило ряди Anthropic і знову зміцнило «крило безпеки» OpenAI. Онлайн-форуми ефективного альтруїзму та безпеки ШІ — головний майданчик, де ці перехресні рухи поширювали, обговорювали й розвивали свої ідеї, — заохочували прихильників іти працювати в провідні лабораторії ШІ, особливо туди, де, на їхню думку, бракувало «вартових безпеки», як-от OpenAI чи DeepMind, щоб впливати на їхню траєкторію розвитку. Наплив нових учасників також сприяв тому, що терміни цієї спільноти стали ширше використовуватися в індустрії ШІ. Те, наскільки швидко, на вашу думку, розвиватиметься ШІ й досягне ключових рубежів, зокрема ШІ (AGI), називають «ШІ-таймлайном». Те, як ви оцінюєте ймовірність того, що ШІ призведе до катастрофічних наслідків, тобто загибелі більшості людства, або до екзистенційних наслідків, тобто повного вимирання людства, — це p(doom), скорочення від «вірогідність загибелі». «Надлишок обчислювальних потужностей» (термін, згаданий у дослідницькій дорожній карті OpenAI 2021 року) та «швидкий перехід штучного інтелекту до надінтелекту» — процес, у якому ШІ стає достатньо потужним, щоб перехитрити людство. «Ризик конкурентного прискорення» означає ризик запуску загостреної конкуренції між компаніями або державами, що

може небезпечно прискорити розвиток ШІ та скоротити часові горизонти його прогресу.

But for a movement that professed independent thinking, EA was swiftly accelerating in the opposite direction. People attracted to its premise were quickly indoctrinated into a broader set of dogmas, propelled by the promise of more opportunities and resources, and an insular social network that played fast and loose with personal and professional boundaries. Within Silicon Valley in particular, EA people largely worked only with other EA people; they largely lived, partied, dated, and slept only with other EA people. Mixed with the tech industry's deep-rooted sexism and the Bay Area's long-standing polyamorous subcultures, its cultlike fervor, manifested in the worst way, could turn into a toxic cauldron of sex, money, and power; it was leading EA to be plagued by growing allegations of sexual harassment and abuse.

Проте, як на рух, що декларував незалежне мислення, ЕА розвивався в протилежному напрямку. Людей, яких приваблювала його ідея, змушували приймати догми, підсилені обіцянками доступу до можливостей і ресурсів та замкненою соціальною мережею, яка завиграшки порушувала особисті й професійні межі. Зокрема, у Кремнієвій долині прихильники ЕА здебільшого працювали лише з іншими членами руху; вони жили, святкували, зустрічалися й вступали в інтимні стосунки переважно всередині спільноти. У поєднанні з глибоко вкоріненим сексизмом техіндустрії та давніми поліаморними субкультурами Затоки, ця культоподібна завзятість у своїх найгірших проявах перетворювалася на токсичну суміш сексу, грошей і влади — і на рух посипалися дедалі чисельніші звинувачення в сексуальних домаганнях та зловживаннях.

In November 2022, SBF's spectacular downfall with the collapse of FTX, along with his sweeping fraud convictions and ensuing twenty-five- year prison sentence, would be to many a symptom of the rot that had festered in the movement. Just as quickly as it caught on, EA fell out of fashion within the tech industry, and

У листопаді 2022 року стрімкий занепад СБФ після краху FTX, масштабні вироки за шахрайство та подальший двадцятип'ятирічний тюремний термін для багатьох стали симптомом гнилі, що давно роз'їдала рух ізсередини. ЕА так само стрімко втратив популярність у техіндустрії, як і здобув її, чимало

many people rapidly disaffiliated.

But even without the label, the movement's social networks, its values and lingo, and the prominence it secured for existential AI safety issues would persist. It would also give rise to a countervailing force: e/acc (pronounced "ee-ack"), or effective accelerationism. What began largely as a joke to lampoon the EA movement would quickly enshrine its polar opposite spirit: Where EA and the broader AI safety community cultivated the most extreme perspectives about slowing down and even slamming the brakes on AI development, or, as in Amodei's view, accelerating AI development while throttling AI adoption, e/acc would elevate the maximalist view of flooring the accelerator on both. For the latter's adherents, technological progress is not just universally good, it's a moral imperative to make that progress as fast as possible. The two groups became colloquially known as the Doomers and Boomers.

Within this bubble, some would begin to view Anthropic and OpenAI as the respective faces of each movement. Others would view OpenAI as a battleground for the polarized ideologies, an organization once rooted in Doomer thinking as a nonprofit that was being yanked away by Boomers with its increasing emphasis, through

людей поспіхом відмежувалися від нього.

Навіть без самої назви рух зберіг свої соціальні мережі, цінності й характерну лексику — так само як і вплив, якого досяг у просуванні екзистенційних питань безпеки ШІ. Водночас він спричинив появу протидіючої течії — ефективного акселераціонізму. Те, що спершу було радше жартом, покликаним висміяти ЕА, швидко оформилося в ідеологію з діаметрально протилежним духом. Якщо ЕА та ширша спільнота з безпеки ШІ культивували найрадикальніші погляди на необхідність сповільнити розвиток ШІ чи, як у випадку Амодея, прискорювати розробку, але стримувати впровадження, то ефективний акселераціонізм підносив до максимуму протилежну позицію: втиснути педаль газу до підлоги і в розробці, і у впровадженні. Для його прихильників технологічний прогрес — це не просто благо, а моральний імператив здійснювати його якомога швидше. У розмовній мові ці два табори почали називати «песимісти» та «прогресисти».

У межах цієї бульбашки дехто почав сприймати Anthropic і OpenAI як відповідні обличчя двох рухів. Інші ж бачили в OpenAI поле бою між поляризованими ідеологіями — організацію, що виникла з «песимістичним» мислення як некомерційний проєкт, але яку поступово перетягували члени «прогресивного» мислення із

its for-profit arm, on making money. Many were uncertain about Altman's allegiance, citing different times he seemed sympathetic to both. Those who were more charitable viewed him as somewhere in the middle, dealing with the tough job of representing all of the different perspectives within his company. But beginning with The Divorce, and the personal fallout between Altman and the Anthropic cofounders, more and more Doomers would begin to view Altman in the worst light possible. So many of the things that put OpenAI on the map and would bring it increasing commercial success had begun as AI safety projects: scaling laws, code generation, reinforcement learning from human feedback, the combination of these three into incredibly compelling large language and then multimodal models. Many Doomers would feel their work was being co-opted and twisted to achieve something directly antithetical to their core values. In their view, it was Altman that was doing that co-opting and twisting. And that made him a pathological liar, a manipulative abuser, and his own threat to humanity.

Soon enough, the clash between these polarized ideologies within OpenAI and its surrounding environment would threaten to tear apart the company that had done more than any other to set the

дедалі більшим акцентом на прибутковість через її комерційне крило. Багато хто не міг визначитися з позицією Альтмана, пригадуючи моменти, коли він, здавалося, симпатизував обом таборам. Ті, хто ставився до нього поблажливіше, вбачали в цьому спробу втримати баланс і представляти різні позиції всередині компанії. Але після «Розлучення» та особистого розриву між Альтманом і співзасновниками Anthropic дедалі більше Песимісти почали дивитися на нього в найгіршому світлі. Чимало з того, що зробило OpenAI впізнаваною і принесло їй комерційний прорив, починалося як проєкти з безпеки ШІ: закони масштабування, генерація коду, навчання з підкріпленням на основі людського зворотного зв'язку та поєднання цих підходів у надзвичайно переконливі великі мовні, а згодом і мультимодальні моделі. Для багатьох Песимістів це виглядало як привласнення й перекручення їхньої роботи задля цілей, які вони вважали прямо протилежними своїм фундаментальним цінностям. У їхньому баченні саме Альтман здійснив це привласнення й перекручення, а отже був патологічним брехуном, маніпулятивним аб'юзером і ще однією загрозою людству.

Невдовзі протистояння цих радикально протилежних ідеологій усередині OpenAI та довкола неї почало загрожувати розірвати компанію, яка більше за будь-яку іншу визначила тон

tone of the new era of AI development. But as much as each ideology professed to be the opposite to the other, both were in fact preaching from the same bible. Both discussed AGI as an increasingly foregone conclusion and with a religious ferocity; both fixated on the long term and asserted a moral authority to keep AI development within the control of its adherents. Where one warned of fire and brimstone, the other tantalized with visions of heaven.

In early 2022, OpenAI was ready to test a different product release strategy, this time with its text-to-image work. It would neither hide the model behind an API nor hand off the product and brand to Microsoft. OpenAI would do the release itself and put the technology directly into the hands of consumers. The model even had an eye-catching name from the original researchers who'd developed it in the company: DALL-E 2, a play off the Spanish surrealist artist Salvador Dalí and the titular robot in the Disney Pixar movie *WALL-E*.

DALL-E had spun out of a trend in the broader field of AI research to develop multimodal models—models that combine at least two different "modalities," such as text, images, sound, or video. For years the field had been working to merge the first two—

нової ери розвитку ШІ. Попри те, що кожна з ідеологій проголошувала себе прямо протилежною іншій, по суті вони ґрунтувалися на одній і тій же «біблії». Обидві сприймали ШІ як дедалі неминучіший результат і з релігійною завзятістю підходили до цієї ідеї; обидві фокусувалися на довгостроковій перспективі та претендували на моральне право тримати розвиток ШІ під контролем своїх прихильників. Там, де одна застерігала «вогнем і сіркою» пекла, інша спокушала видіннями раю.

На початку 2022 року OpenAI була готова випробувати нову стратегію запуску продукту — цього разу щодо своєї текстово-візуальної моделі. Компанія вирішила не ховати її за прикладним програмним інтерфейсом (API) й не передавати продукт і бренд Microsoft. Натомість, OpenAI здійснила реліз самостійно й безпосередньо відкрила технологію для користувачів. Модель отримала гучну назву — *DALL-E 2*, запропоновану її першими розробниками в компанії: гру слів на честь іспанського сюрреаліста Сальвадора Далі та робота WALL-E з мультфільму студії Disney Pixar.

DALL-E з'явилася в руслі ширшого тренду в дослідженнях ШІ — створення мультимодальних моделей, тобто систем, що поєднують щонайменше дві різні «модальності»: текст, зображення, звук або відео. Упродовж років дослідники намагалися об'єднати

language and vision—so a single model would be capable of relating words to visual information. This was driven in part by the data available—text and images are abundant online and the easiest to process—and by a scientific hypothesis: If pure language is not enough to produce human-level intelligence, vision is likely the second most powerful ingredient.

At OpenAI, taking the field as inspiration, the research team had adopted the same progression: After language models, they'd moved on to text-and-image models, and, crucially, focused on continuing to use Transformers in order to retain the model's scalability. While the first Transformer had been initially designed to work best with text, Google had introduced a new Vision Transformer in 2020, adapting it to images.

In January 2021, OpenAI showcased two new Transformer-based models. The first, called CLIP, developed once again by Alec Radford, used the original Transformer and Vision Transformer together to generate detailed captions for images. The second, DALL-E 1, from Aditya Ramesh, a researcher who had studied at New York University and for a time under Meta's Yann LeCun, trained a twelve-billion-parameter Transformer to accept text and generate novel images.

мову й зір, щоб одна модель могла співвідносити слова з візуальною інформацією. Цьому сприяли як практичні обставини, адже тексти й зображення найпоширеніші в Інтернеті й найпростіші для обробки, так і наукова гіпотеза: якщо самих лише мовних даних недостатньо для досягнення людського рівня інтелекту, то зорове сприйняття, ймовірно, є другим за важливістю компонентом.

В OpenAI, наслідуючи загальну логіку розвитку галузі, дослідницька команда пройшла подібний шлях: після мовних моделей вона перейшла до текстово-візуальних і, що принципово, зберегла архітектуру «Трансформер», аби не втратити масштабованість. Хоча перші трансформери спершу створювалися передусім для роботи з текстом, у 2020 році Google представила «Візуальний Трансформер» — адаптовану версію цієї архітектури для обробки зображень.

У січні 2021 року OpenAI представила дві нові моделі на основі архітектури «Трансформер». Перша — *CLIP*, знову розроблена Алеком Редфордом, — поєднувала класичний «Трансформер» і «Візуальний Трансформер», що дозволяло створювати докладні описи зображень. Друга — *DALL-E 1*, створена Адітьєю Рамешем, дослідником із Нью-Йоркського університету, який певний час працював під керівництвом Яна ЛеКуна в Meta, — являла собою «Трансформер» із дванадцятьма мільярдами

In a blog post, OpenAI highlighted DALL-E 1's capabilities with a series of playful prompts, including "an avocado armchair," which produced various green and brown armchairs aesthetically inspired by avocados. The images were slightly blurry and cartoonish, an artifact of the training process that Ramesh had used to produce the model. He had compressed 250 million images to feed them into the Transformer, losing some of their high-resolution details in the process.

As the team started on DALL-E 2, a new method for generating images was gaining traction. Known as diffusion, it was a technique inspired by physics that made it possible for Transformers to better learn the correlations between pixels in a vast swath of images. The original idea had come from a 2015 paper written by Stanford and Berkeley researchers. Five years later, Jonathan Ho, a Berkeley graduate student advised by Pieter Abbeel, one of the early OpenAI researchers, had popularized the technique by cleverly revamping it in ways that generated far more high-fidelity images. Ho also showed that diffusion models could recognize images better than existing computer-vision systems. The findings paralleled Radford's own results with GPT-1: In learning to synthesize

параметрів, навчений приймати текстові запити й генерувати нові, оригінальні зображення.

Усвоєму блозі OpenAI продемонструвала можливості *DALL-E 1* за допомогою серії грайливих запитів, зокрема «крісло-авокадо», який породжував різні зелені й коричневі крісла, естетично натхненні формою фрукта. Зображення були дещо розмитими й мультяшними внаслідок способу навчання, який Рамеш застосував під час створення моделі. Він стиснув 250 мільйонів зображень, щоб завантажити їх у «Трансформер», і в процесі втратив частину високої роздільності.

Коли команда почала працювати над *DALL-E 2*, популярності якраз набирав новий метод генерації зображень — дифузія. Натхненний фізикою підхід дозволяв трансформерам точніше вивчати взаємозв'язки між пікселями у величезних масивах зображень. Першоджерело ідеї - наукова стаття 2015 року, написаній дослідниками зі Стенфорда та Берклі. П'ять років по тому Джонатан Хо — аспірант з Берклі, науковим керівником якого був Пітер Аббіль, один із перших дослідників OpenAI, — переосмислив цей метод і суттєво його вдосконалив, що дозволило отримувати зображення значно вищої якості. Він також показав, що дифузійні моделі можуть розпізнавати зображення краще за тодішні системи комп'ютерного зору. Ці результати перегукувалися з відкриттями

convincing images—the equivalent of generating humanlike sentences—diffusion models had captured the patterns within their training data at a deep enough level to perform a broader range of tasks in visual processing.

OpenAI changed tack to building DALL-E 2 with diffusion and Radford's CLIP. Ramesh and other researchers gradually scaled up the model and added the ability to inpaint—allowing a user to erase a person's hair in a photo and change its color, or select a grassy meadow in a picture and populate it with roaming zebras. Using diffusion created much sharper and more photorealistic images; the method also significantly reduced the amount of compute needed to achieve the same performance as DALL-E 1.

Researchers outside of OpenAI would shrink the compute intensity of diffusion models even further. Stable Diffusion, the popular open-source image generator, would require only 256 Nvidia A100s to train, using a revised technique known as latent diffusion. Björn Ommer, a professor at the Ludwig Maximilian University of Munich whose lab created Stable Diffusion, says he developed the technique after watching image generators go the way of large

Редфорда щодо *GPT-1*: навчившись створювати переконливі зображення — візуальний еквівалент людськоподібних речень, — дифузійні моделі засвоїли закономірності своїх навчальних даних настільки глибоко, що змогли виконувати ширший спектр завдань із візуальної обробки.

OpenAI змінила курс і взялася за розробку *DALL-E 2* на основі дифузії у поєднанні з *CLIP* Редфорда. Рамеш та інші дослідники поступово масштабували модель і додали функцію «inpainting» — можливість редагувати окремі ділянки зображення: наприклад, стерти волосся людини на фото й змінити його колір або виділити трав'яну галявину та «заселити» її зебрами, що вільно блукають. Використання дифузії дало змогу створювати значно чіткіші й реалістичніші зображення; водночас цей підхід помітно скоротив обсяг обчислювальних ресурсів, потрібних для досягнення рівня продуктивності *DALL-E 1*.

Дослідники поза OpenAI зуміли ще більше зменшити обчислювальні витрати дифузійних моделей. «Stable Diffusion» — популярний генератор зображень з відкритим кодом — потребував для навчання лише 256 графічних процесорів Nvidia A100, використовуючи вдосконалений підхід, відомий як латентна дифузія. Бйорн Оммер, професор Мюнхенського університету Людвіга—Максиміліана, чия лабораторія створила «Stable

language models and grow obscenely costly. “We were stuck on a train which was going in the direction of—not just training—but inference actually taking supercomputers to run; millions of dollars of investments,” he says. “We were wondering, could we get the larger research community back in the game and make sure the field of generative AI is not moving in the direction where just a handful of big tech companies would have the required resources to run and to host those models?”

OpenAI wouldn’t adopt latent diffusion until much later, leaving DALL-E 2 and 3 much more computationally expensive than Stable Diffusion or Midjourney, which many users deemed the higher-quality products. It was just one example of how, even within the narrow realm of generative AI, scale was not the only, or even the highest-performing, path to more expanded AI capabilities.

With DALL-E 2’s remarkable jump in performance, the Applied division began working in late 2021 and early 2022 on different ideas for productization. It settled on a web app called Labs that would allow users to play around with the model—and other future models—through a browser. Both product head Fraser Kelton

Diffusion», пояснював, що розробив цей метод, спостерігаючи, як генератори зображень повторюють шлях великих мовних моделей і стають абсурдно дорогими. «Ми опинилися на поїзді, який мчав у напрямку, де не лише навчання, а й сам інференс вимагали суперкомп’ютерів — мільйонних інвестицій», — казав він. — «Тож замислилися: чи можемо ми повернути ширшу дослідницьку спільноту в гру й не допустити, щоб генеративний ШІ рухався до ситуації, коли лише кілька великих техкомпаній матимуть ресурси для запуску й підтримки таких моделей?».

OpenAI перейшла до латентної дифузії значно пізніше, через що *DALL-E 2 і 3* залишалися набагато більш обчислювально витратними, ніж «Stable Diffusion» або «Midjourney», які багато користувачів вважали продуктами вищої якості. Це був лише один із прикладів того, що навіть у відносно вузькій сфері генеративного ШІ масштабування не є ані єдиним, ані обов’язково найрезультативнішим шляхом до розширення можливостей штучного інтелекту.

Після разючого стрибка в можливостях *DALL-E 2* прикладний підрозділ OpenAI наприкінці 2021 — на початку 2022 року почав продумувати різні способи перетворення моделі на продукт. Урешті команда зупинилася на вебзастосунку під назвою Labs, який давав змогу користувачам експериментувати з моделлю

and VP Bob McGrew believed that such an interactive experience would satisfy the clear demand they noticed from GitHub Copilot that people had for engaging directly with generative AI models. It would also help serve the company's mission: DALL-E 2 was fun and delightful, a great way to ease people's fears about powerful AI systems and pave the way for OpenAI to deliver more of its technology's benefits in future releases.

With a still relatively small product staff, the company recruited a few others to help with the website's design and development. To those new members, who hailed from more traditional corporate backgrounds, OpenAI still felt more like working at a university research lab than at a company. Days were often spent reading academic papers and having theoretical debates instead of reviewing mock-ups for interfaces. But to some researchers, the growing presence of Applied staff in their research meetings made them feel the opposite. Gone were the days when all of it was spent on purely exploratory research, like discussing fundamentally new ideas about how to make a better multimodal model; now a growing fraction of their research was in service of

— а згодом і з іншими майбутніми моделями — просто в браузері. І керівник продукту Фрейзер Келтон, і віцепрезидент Боб Макґрю були переконані, що такий інтерактивний формат відповідає очевидному попиту, який вони побачили на прикладі GitHub Copilot: люди хотіли безпосередньо взаємодіяти з генеративними моделями ШІ. Водночас, це добре вписувалося в місію компанії. *DALL-E 2* була веселою й захопливою — зручною точкою входу, що могла зменшити страхи навколо потужних систем ШІ й підготувати ґрунт для подальшого впровадження технологій OpenAI в наступних релізах.

Маючи все ще відносно невелику продуктову команду, компанія залучила кількох додаткових фахівців до розробки й дизайну сайту. Для новачків, які прийшли з більш традиційного корпоративного середовища, OpenAI і надалі більше нагадувала університетську дослідницьку лабораторію, ніж звичайну компанію. Дні часто минали за читанням наукових статей і теоретичними дебатами, а не за обговоренням макетів інтерфейсів. Водночас для частини дослідників зростаюча присутність співробітників прикладного підрозділу на їхніх зустрічах означала протилежне. Минали часи, коли вся увага зосереджувалася на суто дослідницьких експериментах — на кшталт обговорення принципово нових підходів до створення кращих мультимодальних

commercialization, such as figuring out how to optimize existing models for serving up to users.

After the experience of firefighting text-based child sex abuse with AI Dungeon, of particular concern was the possibility of DALL-E 2 being used to manipulate real or create synthetic child sexual abuse material, or CSAM. As with each GPT model, the training data for each subsequent DALL-E model was growing more and more polluted. For DALL-E 2, the research team had signed a licensing deal with stock photo platform Shutterstock and done a massive scrape of Twitter to add to its existing collection of 250 million images. The Twitter dataset in particular was riddled with pornographic content. Several employees made a significant effort to check for and cull any CSAM.

But after some discussion, the employees left in other types of sexual images, in part because they felt such content was part of the human experience. Keeping such photos in the training data, however, meant the model would still be able to produce synthetic CSAM. In the same way DALL-E could generate an avocado armchair having only ever seen avocados and armchairs, DALL-E 2

моделей. Тепер дедалі більша частина роботи підпорядковувалася комерціалізації: наприклад, пошуку способів оптимізувати наявні моделі для їх масштабного використання користувачами.

Після досвіду боротьби з текстовими випадками сексуальної експлуатації дітей у AI Dungeon особливе занепокоєння викликала можливість використання *DALL-E 2* для редагування реальних зображень або створення синтетичних матеріалів сексуального насильства над дітьми (CSAM). Як і у випадку з моделями *GPT*, навчальні дані для кожної нової версії *DALL-E* ставали дедалі більш «зашумленими». У випадку *DALL-E 2* дослідницька команда уклала ліцензійну угоду зі стоковою платформою Shutterstock і здійснила масштабний збір даних з Twitter, додавши ці дані до вже наявної колекції з 250 мільйонів зображень. Особливо проблемною була база двних з Twitter — він містив значний обсяг порнографічного контенту. Кілька співробітників доклали значних зусиль, щоб перевірити й вилучити будь-які матеріали, які могли містити CSAM.

Після обговорення співробітники вирішили залишити інші види сексуальних зображень, частково з огляду на те, що вважали такий контент частиною людського досвіду. Водночас, збереження цих матеріалів у навчальних даних означало, що модель усе ще могла б синтетично відтворювати матеріали сексуального насильства над дітьми (CSAM). Аналогічно до того, як *DALL-E*

and DALL-E 3 could do the same thing with children and porn for child pornography, a capability known as “compositional generation.”

Without filtering the data to address the root of the problem, the burden shifted to building out abuse-prevention mechanisms around the model. This included updated content-moderation filters that wrapped around the model to block abusive images in addition to text as well as a user- behavior-monitoring platform and a so-called ban infrastructure—systems that automatically suspended user accounts that reached a certain threshold of repeat offenses. The company brought on a new head of trust and safety, Dave Willner, who as an early employee at Facebook had written that platform’s very first content standards.

Later, during the development of DALL-E 3, when the data imperative had grown even larger, the research team decided that sexual images were no longer just a “nice to have” but a “need to have.” The share of pornographic images on the internet was so large that removing them shrank the training dataset enough to notably

могла створити «крісло з авокадо», маючи у своєму розпорядженні лише зображення авокадо й крісел, *DALL-E 2* та *DALL-E 3* були здатні поєднувати окремі візуальні категорії — наприклад, дітей і порнографічний контент — у нові зображення. Така здатність називається «композиційною генерацією».

Оскільки дані не фільтрувалися на рівні першопричини проблеми, основний акцент змістився на створення механізмів запобігання зловживанням навколо самої моделі. Це передбачало оновлені фільтри модерації контенту, які працювали поверх моделі й блокували не лише текст, а й небажані зображення, а також систему моніторингу поведінки користувачів і так звану інфраструктуру обмежень — механізми автоматичного призупинення акаунтів після досягнення певного порогу повторних порушень. Компанія також запросила нового керівника з питань довіри та безпеки — Дейва Віллнера, який свого часу розробив перші стандарти модерації контенту для Facebook.

Пізніше, під час розробки *DALL-E 3*, коли потреба у великих обсягах даних стала ще нагальнішою, дослідницька команда вирішила, що зображення сексуального характеру більше не лише «бажані», а фактично «необхідні». Частка порнографічного контенту в Інтернеті виявилася настільки значною, що його

degrade the model's performance. In particular, it made the model worse at generating faces of women and people of color due to the same discovery that Deborah Raji made as a Clarifai intern: A significant share of the online content depicting both groups is sexually explicit. For the same reasons, the researchers left in some other kinds of disturbing images.

In December 2023, an alarmed AI engineer at Microsoft, Shane Jones, would discover the downstream consequences of those decisions. As he played around with Copilot Designer, Microsoft's image generator built on DALL-E 3, he was horrified by how quickly it spit out offensive and sexualized images with little prompting. Just adding the term "pro-choice" into the prompt, Jones found, produced scenes of a demon eating an infant and what appeared to be a drill labeled "pro choice" being used to mutilate a baby. Just prompting the tool for a "car accident" and nothing else produced sexualized women next to violent car crashes, including one in lingerie kneeling by a totaled vehicle, CNBC subsequently found through its own testing.

For three months, Jones petitioned Microsoft executives to

вилучення суттєво зменшувало тренувальний набір даних і помітно погіршувало якість роботи моделі. Зокрема, це призводило до гіршої генерації облич жінок і людей із расових меншин — явища, на яке раніше звертала увагу Дебора Раджі під час стажування в Clarifai: значна частина онлайн-контенту, що зображає обидві ці групи, є сексуалізованою. З тих самих причин дослідники залишили в датасеті й деякі інші типи потенційно неприйнятних зображень.

У грудні 2023 року занепокоєний інженер з AI у Microsoft Шейн Джонс зіткнувся з віддаленими наслідками цих рішень. Експериментуючи з Copilot Designer — генератором зображень Microsoft на базі *DALL-E 3*, — він був вражений тим, як швидко система починала створювати образливі та сексуалізовані зображення навіть за мінімальних підказок. За його словами, додавання до запиту терміна «pro-choice» призводило до появи зображень з демоном, який пожирає немовля, або з об'єктом, схожим на дриль із написом «pro choice», що використовується для завдання шкоди дитині. Навіть простий запит «car accident» без жодних уточнень породжував сексуалізовані образи жінок поруч із жорстокими сценами аварій — зокрема, як згодом встановив CNBC під час власного тестування, зображення жінки в білизні, яка стоїть на колінах біля розбитого автомобіля.

Протягом трьох місяців Джонс звертався до керівництва

take down the tool until it had better guardrails, or at the very least restrict its rating in the Google and Android app store from “E for Everyone” to one for mature audiences. After Microsoft declined to adopt his recommendation and OpenAI was unresponsive, he sent a letter to the Federal Trade Commission. “They have failed to implement these changes and continue to market the product to ‘Anyone. Anywhere. Any Device,’ ” he wrote to the FTC. This problem “has been known by Microsoft and OpenAI prior to the public release of the AI model last October.” Microsoft did not comment on the latest status or outcome of Jones’s letter.

As the launch of DALL-E 2 drew closer, the fighting between OpenAI’s Applied division and the newly restocked Safety clan returned.

For those on Safety, now dispersed across various teams under the Research division, the unprecedented realism of DALL-E 2 brought with it a wide array of unknowns. How could it be weaponized to produce synthetic CSAM or political deepfakes? To manipulate and persuade people? To abuse and harm individuals or create whole-of-society detrimental impacts in other ways that were beyond OpenAI’s foresight and imagination? They urged the

Microsoft із проханням тимчасово вилучити застосунок з доступу, доки не буде запроваджено надійніші запобіжники, або принаймні змінити рейтинг застосунку у магазинах Google і Android із «E for Everyone» на категорію для дорослої аудиторії. Коли Microsoft відмовилася виконати його рекомендацію, а OpenAI не відреагувала, він надіслав листа до Федеральної торговельної комісії США (FTC). «Вони не впровадили цих змін і продовжують просувати продукт як доступний “Anyone. Anywhere. Any Device”», — написав він у зверненні до FTC. За його словами, ця проблема «була відома Microsoft і OpenAI ще до публічного запуску моделі торік у жовтні». Microsoft не прокоментувала актуальний статус чи результати розгляду звернення Джонса.

Із наближенням запуску *DALL-E 2* протистояння між прикладним підрозділом OpenAI та оновленою командою з безпеки відновилося.

Для команди безпеки, які тепер були розпорошені по різних командах у межах дослідницького підрозділу, безпрецедентна правдоподібність *DALL-E 2* принесла з собою широкий спектр невідомих ризиків. Чи може модель бути використана для створення синтетичного CSAM або політичних дипфейків? Для маніпуляції та переконування людей? Для зловживань, завдання шкоди окремим особам або навіть спричинення масштабних негативних наслідків

company not to release the model without further rigorous testing and evidence that it wouldn't produce harm.

For those on Applied, the ever-expanding list of concerns once again seemed hysterical and the bar for release completely unrealistic. No system could ever result in *zero* harm, and certainly not one that stayed in a lab environment and never made contact with real users. Just as Safety worried about the limitations of OpenAI's foresight, Applied believed this was precisely why it needed to release DALL-E 2. Releasing AI models in controlled ways to gain real-world feedback would take away that guesswork and was thus a necessary part of improving their safety.

Central to the clash was an intensifying disagreement over what exactly OpenAI was. To the Safety clan, OpenAI was still an idealistic nonprofit- governed research lab with a paramount obligation to, as stated in its charter, place the benefit of humanity over any commercial interests. Under this premise, the benefits far outweighed the costs of withholding models as long as necessary to

для всього суспільства — таких, які виходять за межі передбачень і уявлень OpenAI? Вони наполягали на тому, щоб компанія не випускала модель без додаткового, суворого тестування й переконливих доказів того, що вона не завдаватиме шкоди.

Натомість для команди прикладної розробки дедалі довший перелік застережень виглядав істеричним, а планка вимог до релізу — цілковито нереалістичною. Жодна система не може гарантувати нульовий ризик шкоди — і вже точно не та, що залишається в лабораторії та ніколи не стикається з реальними користувачами. Якщо команда безпеки хвилювалася через обмеженість передбачень OpenAI, то команда прикладної розробки вважала, що саме це й є підставою для випуску *DALL-E 2*. На їхню думку, лише контрольований реліз моделей ШІ з подальшим збиранням зворотного зв'язку в реальному середовищі дозволяє зменшити припущення й невизначеність — і, зрештою, підвищити їхню безпеку.

У центрі цього протистояння було дедалі гостріше розходження в розумінні того, чим насправді є OpenAI. Для команди безпеки це залишалася ідеалістична дослідницька лабораторія під управлінням некомерційної структури, з головним зобов'язанням — як зазначено в її статуті — ставити благо людства вище за будь-які комерційні інтереси. За такої логіки вигоди від відтермінування

think through as many downsides as possible and research ways to mitigate them. To Applied, OpenAI needed to make more practical decisions, grounded in the realities of how the world worked. Essential to the company's mission was remaining a leader in AI research to establish norms around the technology's development. That meant tolerating a degree of risk to move quickly, especially with rumblings of Google finalizing its own image generator, as well as securing the extraordinary capital needed to continue doing cutting-edge research. The latter required raising money from investors, which required working in good faith to advance a commercial strategy that would one day provide those investors returns.

The people in Safety were "completely naive" about the way companies, and the world, work, says a former employee in Applied. "Well, the stakes of OpenAI's proposed AGI mission are high," says another in Safety. " 'Normal company' maybe isn't good enough."

Different teams were codifying this growing conflict into the metrics they used to evaluate their performance. Within the Applied division, the product team and a budding go-to-market operation

релізу моделей — настільки довго, наскільки потрібно, щоб ретельно зважити потенційні ризики й дослідити способи їх мінімізації, — суттєво переважали можливі втрати. Для команди прикладної розробки же OpenAI мала діяти прагматичніше, з огляду на реальні умови ринку. Важливою частиною її місії було зберігати лідерство в дослідженнях ШІ та формувати норми розвитку цієї технології. Це означало готовність прийняти певний рівень ризику заради швидкого руху вперед — особливо на тлі повідомлень про те, що Google готує власний генератор зображень, — а також потребу залучати колосальні інвестиції для підтримки передових досліджень. А це, своєю чергою, вимагало співпраці з інвесторами й добросовісного просування комерційної стратегії, яка з часом могла б забезпечити їм повернення вкладених коштів.

Один із колишніх співробітників команди прикладної розробки назвав підхід команди безпеки щодо того, як насправді працюють компанії і світ загалом, — «цілковито наївним». «Але ставки місії OpenAI зі створення AGI надзвичайно високі, — відповідає інший представник команди безпеки. — Можливо, логіки "звичайної компанії" тут просто недостатньо».

Різні команди почали відображати цей дедалі глибший конфлікт у показниках, за якими оцінювали свою роботу. У підрозділі прикладної розробки команда продукт-менеджерів разом

were developing user growth and revenue targets. Within the Research division, the various AI safety teams struggled to find quantifiable ways of measuring their advancement when it was difficult to specify by nature. AI safety was still a comparatively young discipline. There were no obvious and established benchmarks. In meetings and on Slack, people in Safety repeatedly raised concerns to senior leadership about how this imbalance was causing misaligned incentives: Having clear-cut growth and revenue goals without some kind of strong, comparable counterbalance was pushing OpenAI to operate more and more like a “move fast and break things” operation.

In private conversations with Safety, Altman expressed sympathy for their perspective, agreeing that the company was not on track with its AI safety research and needed to invest in it more. In private conversations with Applied, he pressed them to keep going. During board meetings, he nodded along as Brockman voiced frustrations about the ways that people were using AI safety as political leverage to stall progress for their own purposes.

із новоствореним напрямом go-to-market встановлювала цілі зі зростання кількості користувачів і доходів. Натомість у дослідницькому підрозділі команди, що працювали над безпекою ШІ, намагалися знайти вимірювані способи оцінити власний прогрес — завдання, яке за своєю природою важко реалізувати. Безпека ШІ залишалася відносно молодою галуззю, у якій ще не було чітких і загальновизнаних критеріїв оцінювання. Не було очевидних і встановлених орієнтирів. На нарадах і в Slack співробітники відділу безпеки неодноразово висловлювали керівництву занепокоєння щодо того, як цей розподіл завдає шкоди: цілі зі зростання й доходів не мали аналогічної противаги, OpenAI дедалі більше починала працювати за принципом «рухайся швидко й ламай усе».

У приватних розмовах із представниками команди безпеки Альтман висловлював співчуття до їхньої позиції, погоджуючись, що компанія відстає у сфері досліджень безпеки ШІ й має більше в них інвестувати. У приватних розмовах із командою прикладної розробки він, навпаки, підштовхував їх не зупинятися й рухатися далі. На засіданнях ради директорів він кивав на знак згоди, коли Брокман висловлював роздратування тим, що безпеку ШІ використовують як політичний інструмент для гальмування прогресу з власних міркувань.

More and more, Mira Murati played the role of negotiator, smoothing out the fault lines between different factions and searching for ways to thread the needle between them. On DALL-E 2, she struck a compromise: The web app would be released not as a product but as a “low-key research preview.” Such branding would give OpenAI more leeway to place harsher restrictions on the model, satisfying Safety, while still giving the company a chance to trial a direct-to-consumer relationship and gather user feedback, pleasing Applied. It was also a practical measure. OpenAI didn’t yet have the infrastructure in place for content moderating generated images. Calling the model a “research preview” and not charging for it would allow the company to use blunt, overly broad blockers without fear of upsetting paid users, to buy time for developing more sophisticated filters. The company moved forward with implementing a series of aggressive abuse-prevention mechanisms, including disabling DALL-E 2’s ability to generate any photorealistic faces or edit any real photos with faces to completely circumvent the synthetic CSAM and political misinformation problem.

In March 2022, OpenAI released DALL-E 2 via the Labs web

З часом Міра Мураті дедалі частіше брала на себе роль посередниці, згладжуючи лінії розлому між різними таборами й намагаючись знайти рішення, яке дозволило б віднайти баланс. Щодо *DALL-E 2* вона запропонувала компроміс: вебзастосунок мав вийти не як повноцінний продукт, а як «research preview» — стриманий дослідницький реліз. Таке позиціонування давало OpenAI більше простору для запровадження суворіших обмежень, чим задовольняло команду безпеки, і водночас дозволяло протестувати модель у прямій взаємодії з користувачами та зібрати зворотний зв’язок, чого прагнула команда прикладної розробки. Це було й практичне рішення: на той момент OpenAI ще не мала належної інфраструктури для модерації згенерованих зображень. Називаючи модель «дослідницьким релізом» і не запроваджуючи оплату, компанія могла використовувати грубі, надмірно широкі блокування, не ризикуючи роздратувати платних користувачів, і таким чином виграти час для розробки більш точних фільтрів. Компанія також почала впроваджувати низку жорстких механізмів запобігання зловживанням, зокрема повністю вимкнула здатність *DALL-E 2* генерувати будь-які фотореалістичні обличчя або редагувати реальні фотографії з обличчями, щоб обійти проблему синтетичного CSAM і політичної дезінформації.

У березні 2022 року OpenAI запустила *DALL-E 2* через

app to overwhelming public enthusiasm. As people gushed over and grappled with the model's capabilities, to a degree that exceeded many employees' expectations, the web app went viral across social media, producing a plethora of wild, wacky, and surreal AI-generated art in its wake. It was a GPT-3 moment but better. Instead of engaging with only a small pool of technical developers, the company was tapping into a much broader and more global base of consumers. In real time, it could also respond to user feedback with instantaneous changes to the Labs web app. " This is intoxicating," Fraser Kelton would remember of the experience in a podcast.

Over the next few months, the Applied division, which hadn't yet thought much at all about how to monetize DALL-E 2, raced to turn the web app into a paid offering. It worked with artists and creative professionals around the world to incorporate DALL-E 2 into their practice.

It rolled out a beta program, inviting one million people around the world to get access to the model with free credits for image generations. But as OpenAI started charging, it wasn't Google that proved to be the main challenger, though the tech giant did indeed follow quickly with its Imagen model. Instead, it was two models

вебзастосунок Labs — і реакція публіки виявилася приголомшливою. Користувачі з ентузіазмом експериментували з можливостями моделі — навіть активніше, ніж очікувалося. Labs стрімко став вірусним у соцмережах, породивши хвилю химерного, дотепного й сюрреалістичного ШІ-мистецтва. Це був момент масштабу *GPT-3* — але ще потужніший. Тепер компанія взаємодіяла не лише з обмеженим колом технічних розробників, а з набагато ширшою й глобальнішою аудиторією користувачів. До того ж, вона могла в реальному часі реагувати на відгуки, миттєво вносячи зміни до Labs. «Це п'янить», — згадував згодом Фрейзер Келтон у подкасті.

Упродовж наступних кількох місяців підрозділ Applied, який раніше майже не замислювався над монетизацією *DALL-E 2*, поспіхом узявся перетворювати вебзастосунок на платний сервіс. Команда співпрацювала з художниками й креативними професіоналами з усього світу, допомагаючи їм інтегрувати *DALL-E 2* у свою роботу.

OpenAI запустила бета-програму, запросивши близько мільйона людей у всьому світі й надавши їм безкоштовні кредити на генерацію зображень. Однак щойно компанія почала стягувати оплату, головним конкурентом виявилася не Google — хоча техногігант і справді швидко представив власну модель «Imagen».

from startups, Midjourney and Stability AI's Stable Diffusion. Both image generators were free to use and just as good, if not better, than DALL-E 2 and had fewer safety measures, including allowing users to generate and edit faces, even of politicians. As DALL-E 2 rapidly lost traction in the market, the experience left Applied with a nagging sense that it had lost out on a major commercial opportunity due to, among other things, the app being too restrictive. The team had already been in the process of unwinding its blunt blockers and replacing them with more targeted guardrails. Fueled by a desire to outrace competitors, executives were now pushing the team to unwind them as fast as possible.

To lift the ban on faces, OpenAI developed a new process for preventing and cracking down on the generation of harmful images of people, including CSAM. It used automated systems to detect when faces were being generated in acceptable or abusive contexts and once again relied on overseas contractors to help with the content moderation. This time those contractors were based in India through a vendor called Cogito and reviewed not just reams of text but images—synthetic and real—of the kinds of sexual and violent content that had been sent to Sama workers. As they sifted through what could be hundreds of images a day, the contractors struggled to

Натомість серйозну загрозу становили два стартапи: «Midjourney» та «Stability AI» зі своїм «Stable Diffusion». Обидва генератори були безкоштовними, не поступалися *DALL-E 2* за якістю — а подекуди й перевершували її — і водночас мали значно м'якші обмеження безпеки, дозволяючи, зокрема, генерувати й редагувати обличчя, навіть політиків. Коли *DALL-E 2* стрімко почала втрачати позиції на ринку, в Applied з'явилося гнітюче відчуття, що компанія втратила значну комерційну можливість — зокрема через надмірну суворість застосунку. Команда вже поступово відмовлялася від грубих блокувань, замінюючи їх точковішими запобіжниками. Але тепер, підживлені прагненням випередити конкурентів, керівники тиснули на команду, вимагаючи знімати ці обмеження якнайшвидше.

Щоб зняти заборону на генерацію облич, OpenAI розробила нову систему для запобігання створенню шкідливих зображень людей і боротьби з ними, зокрема матеріалами CSAM. Було впроваджено автоматизовані інструменти, які визначали, чи з'являються обличчя в припустимих або образливих контекстах, а для модерації знову залучили закордонних підрядників — цього разу працівників в Індії, найнятих через підрядника Cogito. Вони переглядали не лише великі обсяги тексту, а й зображення — синтетичні та реальні — із сексуальним або насильницьким змістом, подібні до тих, із якими раніше працювали модератори Sama.

distinguish between sexual content involving seventeen-year- old minors versus eighteen-year-old legal adults. They also couldn't always tell whether the images were fake or real.

What had, on the face of it, been OpenAI's easiest goal in its 2021 research road map turned out to be one of the hardest: scaling up GPT-3 by 10x with Microsoft's new eighteen thousand Nvidia A100 supercomputer cluster, in its effort to develop what would become GPT-4. One-third of the GPT-3 scaling team had left with The Divorce, taking with them significant technical and institutional knowledge. More existentially, OpenAI had run out of data.

After GPT-3, researchers had sought to accumulate as much data as possible, building up the company's reservoir by downloading every new data dump and scraping every new online forum they stumbled upon that didn't have clear warnings against doing so. And yet, even with the additions of GitHub's large repository for Codex, and the coding textbooks and manuals, it was still not enough.

Переглядаючи іноді сотні зображень на день, підрядники часто не могли впевнено відрізнити сексуальний контент із сімнадцятирічними неповнолітніми від контенту з вісімнадцятирічними повнолітніми, а також визначити, чи є зображення реальним чи згенерованим.

Те, що на перший погляд здавалося найпростішим пунктом у дослідницькій дорожній карті OpenAI на 2021 рік, виявилося одним із найскладніших: масштабувати *GPT-3* у десять разів, використовуючи новий суперкомп'ютерний кластер Microsoft із вісімнадцяти тисяч графічних процесорів Nvidia A100, у межах розробки того, що згодом стане *GPT-4*. Після «Розлучення» компанію залишила третина команди, яка працювала над масштабуванням *GPT-3*, забравши з собою значний обсяг технічних знань та інституційної пам'яті. Була й ще глибша проблема: у OpenAI закінчилися дані.

Після *GPT-3* дослідники намагалися зібрати якомога більше даних, розширюючи «резервуар» компанії: завантажували новий великий масив даних й аналізували кожен новий онлайн-форум, на який натрапляли, якщо там не було чітких застережень проти такого збору даних. І все ж навіть додавання масштабного репозиторію GitHub для Codex, а також підручників і довідників із програмування виявилося недостатнім.

With an uphill battle ahead, the situation had all the characteristics of a Greg Brockman project. Not only would it channel his scrappy can-do attitude and his coding brilliance, but it would also focus his energy, for the sake of the rest of the company, on something productive.

After Altman took over, relieving Brockman of his managerial responsibilities, Brockman had eventually gone back to being an individual contributor with no reports. Yet as the nominal president and one of OpenAI's cofounders, he maintained incredible influence over employees and the strategic direction of the company. As OpenAI professionalized and implemented more standard corporate processes, moving away from the freewheeling days of an early-stage startup, Brockman's mix of low responsibility and high authority turned into a liability.

Just like his college and Stripe days, he was not one for institutions and process. He had a restless and obsessive energy. He rarely attended meetings, and set his own schedule, often preferring to code for dozens of hours straight with few breaks for meals and sleep. With the right project, the effects were miraculous: His intense productivity would supercharge progress. But left idle, he tended to

Зважаючи на складність завдання, ситуація мала всі ознаки проекту Грега Брокмана. Він вимагав і його невтомного «зробимо це» підходу, і технічної майстерності в програмуванні. А ще, на благо всієї компанії, зосереджував свою енергію на чомусь конкретному й продуктивному.

Після того, як Альтман перебрав на себе керівництво і звільнив Брокмана від управлінських обов'язків, той зрештою повернувся до ролі індивідуального виконавця без підлеглих. Водночас, залишаючись номінальним президентом і співзасновником OpenAI, він зберігав величезний вплив як на співробітників, так і на стратегічний курс компанії. У міру того, як OpenAI професіоналізувалася й запроваджувала більш стандартні корпоративні процеси, відходячи від вільнодумної атмосфери раннього стартапу, поєднання низької формальної відповідальності з високим рівнем влади дедалі більше ставало для компанії проблемою.

Як і за часів навчання та роботи у Stripe, він не був людиною інституцій і процедур. Йому була властива невгамовна, майже нав'язлива енергія. Він рідко з'являвся на зустрічах, сам визначав свій розклад і часто міг програмувати десятки годин поспіль, майже без перерв на їжу та сон. Коли в нього був потрібний йому проєкт, ефект здавався майже дивовижним: його шалена продуктивність

create a trail of destruction, popping up in projects all over the place to meddle with and derail long-standing plans with last-minute changes. At times, when employees put up resistance, he would deliver emotional pleas higher and higher up their leadership chain to get what he wanted.

Brockman usually did get what he wanted. Much to the frustration and confusion of other executives, Altman was strangely permissive of his behavior. Not only that, Brockman could also influence Altman into meddling and derailing things for him, if only, it seemed, to satisfy Brockman. One popular guess as to why: Though Altman was Brockman's boss as the CEO, Brockman also had authority over Altman as a board member. It was a strange tangle of a structure that ultimately left nothing and no one to hold Brockman accountable.

The senior leadership had changed his role, scope, and reporting lines several times in an attempt to find the best place for him. As with so much else, the buck eventually passed to Murati, who became Brockman's manager. When she sought to give him feedback, he seemed receptive, but on points where he disagreed, he

стрімко пришвидшувала прогрес. Втім без чіткого завдання, він часто залишав по собі руйнівний слід — втручався в різні проєкти й зривав давно вибудувані плани раптовими змінами в останній момент. Інколи, стикаючись із опором, він переходив до емоційних апеляцій, звертаючись дедалі вище по управлінській вертикалі, щоб отримати бажане.

Зазвичай Брокман таки домагався свого. На подив інших керівників, Альтман дивним чином виявляв до його поведінки значну поблажливість. Більше того, Брокман міг впливати на Альтмана так, що той сам втручався й зривав процеси — здавалося, лише для того, щоб задовольнити його вимоги. Одне з найпоширеніших пояснень полягало в особливостях управлінської конструкції: хоча Альтман як CEO формально був керівником, Брокман водночас мав владу над Альтманом як член ради директорів. Цей дивний вузол повноважень у підсумку призвів до того, що фактично ніхто й ніщо не могло притягнути Брокмана до відповідальності.

Керівництво неодноразово змінювало його роль, сферу відповідальності та лінії підпорядкування, намагаючись знайти оптимальне місце. Як і з багатьма іншими питаннями, ця проблема зрештою перейшла до Мурати, яка стала його безпосередньою керівницею. Коли вона намагалася давати йому зауваження щодо

complained to Altman. Murati slowly gave up on attempting to change things with feedback, instead spending significant time trying to find projects for Brockman where he could be net beneficial rather than chaotic, and, with McGrew, healing the ruptures Brockman caused in various parts of the company.

With roadblocks that needed to be punched through in the way of GPT- 4's development, the stars aligned.

To solve OpenAI's data bottleneck, Brockman turned to a new source: YouTube. OpenAI had previously avoided this option—scraping YouTube to train OpenAI's models, YouTube's CEO would later confirm, violated the platform's terms of service. But under the new existential pressure for more data, the question became whether YouTube, or its parent, Google, would enforce it. If Google cracked down, it could jeopardize its own ability to scrape other websites for its large language model development. Brockman was willing to take the risk.

With a small team, Brockman began collecting YouTube

роботи, він демонстрував готовність його приймати, але в тих моментах, де не погоджувався, звертався зі скаргами до Альтмана. Зрештою Мурата поступово відмовилася від спроб вплинути на ситуацію через обговорення його роботи та натомість витрачала значну частину часу на пошук проєктів, у яких внесок Брокмана був би радше конструктивним, ніж дестабілізуючим, а також — разом із Макґрю — на усунення наслідків його втручань у різних підрозділах компанії.

І коли на шляху розробки GPT-4 з'явилися бар'єри, які потрібно було буквально проламати, обставини склалися для нього якнайкраще.

Для подолання дефіциту даних Брокман звернувся до нового джерела — YouTube. Раніше OpenAI свідомо уникав цього варіанта: збирання даних із YouTube для навчання моделей, як згодом підтвердив CEO платформи, порушувало її умови користування. Але в умовах нового екзистенційного тиску питання полягало в іншому — чи наважаться YouTube або його материнська компанія Google застосувати ці правила на практиці. Якщо Google вдалася б до жорстких заходів, це могло б поставити під загрозу її власну можливість здійснювати вебскрейпінг інших сайтів для розвитку великих мовних моделей. Брокман був готовий піти на цей ризик.

Працюючи з невеликою командою, Брокман почав збирати

videos, eventually compiling more than one million hours of footage, according to *The New York Times*. He then used a speech-recognition tool called Whisper, which Radford had developed, to transcribe the videos into text for GPT-4.

Next was the training. To train GPT-3, the Nest team had designed a bespoke software platform. With most of its creators now gone to Anthropic, they were no longer around to explain how it worked. As a point of pride, some leadership didn't want to rely on the Anthropic team's legacy either. Brockman disappeared into his coding hole and developed a new platform. Then, with several others, including Jakub Pachocki and Szymon Sidor, the Polish scientists whom he'd grown close with during the *Dota 2* project, Brockman babysat GPT-4's training. The pre-training alone took three months.

At first, GPT-4 seemed like a disappointment. "It was a wild model, which in some sense behaved quite poorly," one researcher says. "Because the average data quality was so horrible, and because the model was quite powerful and context sensitive, it was producing garbage responses." But Brockman pushed forward, pulling together

відео з YouTube і, за даними The New York Times, зрештою накопичив понад мільйон годин матеріалу. Для перетворення цього масиву на текст він використав Whisper — інструмент розпізнавання мовлення, розроблений Радфордом, — що транскрибував відео й застосував отримані тексти для навчання GPT-4.

Далі почався етап навчання. Для GPT-3 команда Nest розробила власну програмну платформу. Однак, після «Розлучення» більшість її авторів перейшли до Anthropic, тож було нікому детально пояснити, як вона працює. До того ж частина керівництва з міркувань принципу не хотіла покладатися на технологічну спадщину Anthropic. Брокман знову зник у своєму «кодовому бункері» й створив нову платформу. Потім разом із кількома колегами, зокрема Якубом Пахоцьким і Шимоном Сидором, польськими науковцями, з якими він зблизився під час проєкту Dota 2, він супроводжував весь процес навчання GPT-4. Лише попереднє навчання тривало три місяці.

Спершу GPT-4 радше розчаровувала. «Це була дика модель, яка в певному сенсі поводилася досить погано», — згадує один із дослідників. — «Через украй низьку середню якість даних, а також через її потужність і високу чутливість до контексту, вона часто генерувала безглузді відповіді». Але Брокман не зупинявся,

the resources to improve the model with human contractors conducting reinforcement learning from human feedback. With each week, the results looked better and better, until the performance truly began to wow people internally.

GPT-4 now had built-in multimodal capabilities and, against OpenAI's internal assessments, was generating more polished code than ever and was more nimble in recognizing user intent and delivering helpful answers. In an impressive showcase of those abilities, Brockman would later live stream a demo of him prompting GPT-4 with a photo of a simple chicken scratch sketch of a web page drawn in his notebook. "My Joke Website," Brockman had written at the top. Stacked below it, he'd added: "[really funny joke!]" and "[push to reveal punchline]." In less than half a minute, the model would turn that sketch into workable code, stylizing the first line as a title, replacing the second line with a joke, and recognizing the third line as a button.

But as OpenAI began teasing the model in trusted circles, including investors and select customers, at least one person wasn't the least bit impressed. It was once again the ever-hard-to-please Bill Gates.

In June 2022, after getting a demo of GPT-4, Gates expressed disappointment in the insufficient progress from GPT-2. Despite the

мобілізувавши ресурси для вдосконалення моделі за допомогою навчання з підкріпленням на основі людського зворотного зв'язку. Тиждень за тижнем результати ставали дедалі кращими, доки внутрішні оцінки не почали по-справжньому вражати команду.

GPT-4 уже мала вбудовані мультимодальні можливості й, усупереч внутрішнім очікуванням OpenAI, генерувала більш витончений код, ніж будь-коли раніше, а також значно краще розпізнавала наміри користувачів і надавала доречні відповіді. Вражаючою демонстрацією цих можливостей став пізніший прямий ефір, у якому Брокман показав, як він подає GPT-4 фотографію простого, нашвидкуруч накиданого ескізу вебсторінки у своєму блокноті. Угорі він написав: «Сторінка жартів». Нижче додав: «[смішний жарт!]» та «[натиснути щоб дізнатись жарт]». Менш ніж за пів хвилини модель перетворила цей ескіз на робочий код: оформила перший рядок як заголовок, замінила другий на жарт і правильно інтерпретувала третій як кнопку.

Коли OpenAI почала обережно презентувати модель у вузьких, довірених колах — серед інвесторів і вибраних клієнтів — принаймні одна людина зовсім не була вражена - вічно незадоволений Білл Гейтс.

У червні 2022 року, після демонстрації GPT-4, Ґейтс висловив розчарування прогресом порівняно з GPT-2, який, на його

model being significantly larger and more fluent, he still felt like it was “ an idiot savant,” unable to tackle complex scientific problems. He told the team that he would only start paying attention once GPT-4 scored a 5 on an AP Biology test— AP Bio because he felt it tested critical scientific thinking rather than a memorization of facts. “I thought, ‘Okay, that’ll give me three years to work on HIV and malaria,’ ” Gates later recounted in his podcast.

Brockman took Gates’s remark as a challenge. He immediately reached out to Sal Khan, the CEO of online education platform Khan Academy, and asked him to tap into the company’s large repository of AP Bio questions as training data. Khan was skeptical but agreed to do so in exchange for his platform getting access to the model. Brockman also amassed a team of employees to build a special user interface for the new Gates Demo.

By late August, much to Gates’s surprise, Altman and Brockman were pinging him again. Over dinner at the Microsoft founder’s house the following month with roughly thirty people, the two OpenAI executives and others showed Gates a series of highly refined GPT-4 demos designed to impress him. The crowning

думку, був недостатнім. Хоч модель і стала значно більшою та плавнішою, він усе ще сприймав її як «ідіота-всезнайку», нездатного розв’язувати складні наукові задачі. Він сказав команді, що почне серйозно ставитися до моделі лише тоді, коли вона отримає 5 балів на іспиті AP Biology — саме цей предмет, на його переконання, перевіряє не механічне запам’ятовування фактів, а критичне наукове мислення. «Я подумав: гаразд, це дає мені ще три роки, щоб займатися ВІЛ і малярією», — згадував Ґейтс згодом у своєму подкасті.

Брокман сприйняв слова Ґейтса як виклик. Він одразу зв’язався із Салом Ханом, генеральним директором освітньої онлайн-платформи Khan Academy, і попросив надати доступ до великого масиву завдань з AP Biology для використання в якості навчальних даних. Хан поставився до цієї ідеї скептично, але зрештою погодився — в обмін на доступ його платформи до моделі. Паралельно Брокман зібрав команду, щоб створити спеціальний користувацький інтерфейс для нової демонстрації для Ґейтса.

Наприкінці серпня, на велике здивування Ґейтса, Альтман і Брокман знову написали йому. Наступного місяця, під час вечері в домі засновника Microsoft, де зібралося близько тридцяти гостей, двоє керівників OpenAI разом з іншими учасниками представили Ґейтсу серію спеціально підготовлених, ретельно відшліфованих

moment was the model acing AP Bio: It nailed fifty-nine out of sixty multiple-choice questions and generated impressive answers to six open-ended ones. An outside expert would score the test: 5 out of 5. Gates couldn't believe it. His shock and praise, which the demo attendees would instantly relay back to the rest of the company, ripped like wildfire through the office and incited an exhilarating level of energy: This showcase, Gates said, was one of the two most stunning demos he'd ever seen in his life.

In all-hands meetings, Altman continued to stoke the excitement. "Startups that do remarkable things require a miracle," he said. "We just had our miracle." Many employees believed it, awestruck by the momentousness of what they had accomplished. GPT-4's new level of performance convinced OpenAI leadership that it was time to start working toward one of Altman's long-coveted ambitions: an AI assistant that would look and feel like the character Samantha in the 2013 Spike Jonze movie *Her*.

For years, *Her* had been a touchstone that Altman and other OpenAI cofounders frequently invoked as an example of what AGI might one day look like: a single multimodal model whose product interface felt so utterly natural that it faded away and simply brought

демонстрацій GPT-4. Кульмінацією став результат з AP Biology: модель правильно відповіла на п'ятдесят дев'ять із шістдесяти тестових запитань і згенерувала переконливі відповіді на шість відкритих. Незалежний експерт оцінив роботу — 5 із 5. Ґейтс не міг повірити побаченому. Його подив і схвальні слова, які учасники демонстрації миттєво передали решті компанії, розлетілися офісом, мов лісова пожежа, й спричинили хвилю піднесення. За словами Ґейтса, це була одна з двох найвражаючих демонстрацій, які він бачив у своєму житті.

Під час загальних зустрічей Альтман і далі підігрівав ентузіазм. «Стартапам, які роблять щось по-справжньому визначне, потрібне диво, — сказав він. — І ми щойно його пережили». Багато співробітників повірили в це, приголомшені масштабом досягнутого. Новий рівень можливостей GPT-4 переконав керівництво OpenAI, що настав час рухатися до однієї з давніх амбіцій Альтмана — створення AI-асистента, який виглядатиме й відчуватиметься як Саманта з фільму Her (2013) режисера Спайка Джонза.

Упродовж років «Her» залишався для Альтмана та інших співзасновників OpenAI своєрідним орієнтиром — прикладом того, яким одного дня може стати AGI: єдина мультимодальна модель з інтерфейсом настільки природним, що ніби зникає, залишаючи

user delight. “I would think it’s because it was an assistant that was wonderfully integrated into a life,” says a former employee, of why the movie was such a pivotal reference. “The positive arc of that story before it unravels is a really great story of AI’s evolution into society.”

лише радість від користування. «Думаю, річ у тім, що це був асистент, який органічно вплітався в повсякденне життя», — каже колишній співробітник, пояснюючи, чому фільм став таким джерелом натхнення. — «Позитивна траєкторія цієї історії — до того, як усе починає розпадатися, — це дуже переконливий сценарій інтеграції ШІ у суспільство».

Chapter 2. Peculiarities of translating technical terminology relating to Artificial Intelligence field into Ukrainian

2.1 Genre and thematic characteristics of the book *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI* by Karen Hao

Karen Hao is an American journalist and author known for her reporting on artificial intelligence and its social impact, as well as for her investigative nonfiction work. She began her career with a technical background, earning a Bachelor of Science degree in Mechanical Engineering from the Massachusetts Institute of Technology, before shifting her focus to journalism. Her professional experience includes work at major international publications. In her research and writing, Hao explores the impact of artificial intelligence on society and the environment, as well as issues related to labour and resource use in the technology sector. In addition to her written work, she has contributed to improving public understanding of artificial intelligence through other media. She co-produced the podcast *In Machines We Trust* and founded the newsletter *The Algorithm*, both of which aim to explain AI development to a broad audience.

In 2025, she published the book *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*, which became a New York Times bestseller. The book examines the rise of OpenAI and the global consequences of the development of artificial intelligence. It is based on extensive research, including hundreds of interviews, archival materials, and years of journalistic work, enabling the author to present a detailed, fact-based account.

The central idea of *Empire of AI* is that human decisions, misunderstandings, political and economic conditions, and the influence of key individuals shape the development and perception of the AI industry. Hao argues that OpenAI's original mission — to develop AI for the benefit of humanity — has gradually shifted toward profit-oriented strategies, reduced transparency, safety concerns, and increased competition. This transformation reflects broader trends in the technology sector, where even non-profit organizations may become driven by financial incentives.

The book consists of prologue, four parts composed of 18 chapters, and an epilogue. It begins with the November 2023 crisis over the dismissal and reinstatement of Sam Altman as CEO. It is at this very moment that a chain of decisions and actions begins that will transform the company into a commercial project. The book also addresses the global consequences of AI development, including:

- the exploitation of labour, particularly low-paid data workers,
- large-scale data collection without explicit consent,
- environmental impacts, especially high energy consumption,
- and the concentration of power among a small number of technology companies.

The narrative is primarily presented in the third person, but also incorporates first-person perspectives through quotes and commentary from individuals closely involved in the events.

During translation, it became evident that the book employs complex sentence structures and extensive technical terminology related to artificial intelligence. The frequent references to numerous people, companies, projects, and internal code names. Therefore, particular attention was to be paid to accurately conveying long, information-dense sentences and maintaining clarity in the sequence of events. However, the most difficult challenge in this section of the text was identifying and translating the technical terms.

2.2 Classification of technical terms and the challenges of their translation

Technical terminology in the field of artificial intelligence development posed the greatest challenge during the translation process. We will explore methods of identification and

classification in this section. The first thing to understand before we start analysing is what the concept of the *term* covers. According to the *Handbook of Terminology*, the fundamental principle of terminology can be understood as follows: terms function within specific fields of activity, each of which is organised as a structured classification system of specialized knowledge. Consequently, each field of specialisation has its own system, and this structure must be consistently reflected in any comprehensive terminology collection (Pavel & Nolet 2001, p. 1).

However, while the terminology theoretically has a structure and is linked to specific systems, in practice, these systems are not strictly separate and often overlap, making their meaning unclear and difficult to define. It is difficult to draw a clear boundary between the terminology of related fields, just as it is challenging to distinguish between specialized terms and the vocabulary of the common language. Specialized terms can become part of everyday language; there is a constant flow between them (Cabré & Sager 1999, p. 80). This crossing of specialized and common language is further complicated by the fact that some lexical items may appear to be non-specialized, even though they actually function as precise technical terms within a specific field. After overcoming the challenges presented by specialized terminology, translators must also deal with terms that, at first glance, appear to be part of common language but actually have a very specific meaning within a particular field. Such seemingly ordinary words can be misinterpreted, as their meaning depends largely on the context and subject area (Byrne 1988, p. 52).

Although technical terminology makes up only 5-10% of the text, it is most often used to determine whether a translation is technical (Newmark 1988, p. 151). Despite its relatively small proportion, its role in translation remains important. As Byrne points out, while terminology is considered one of the least problematic aspects of translation, relying solely on the translator's research skills is not always sufficient. Companies tend to use industry-specific terminology in corporate documentation (2014 p. 144). This is particularly relevant for texts such as *Empire of Artificial Intelligence: Dreams and Nightmares in Sam Altman's OpenAI*, which draws on corporate materials related to OpenAI.

Towards the end of the 20th century, humanity entered the era of the information revolution, characterized by global computerization and rapid technological change. This led to a significant expansion of the lexicon, particularly in the English language, a phenomenon often referred to as the 'neologism boom' (Syndega & Ivashchyshyn 2009 p. 352). Information technology (IT) is a specialized field in which new terms seem to appear almost every day. It is therefore reasonable to assume that the number of new IT-related terms may exceed the number of terms associated with all other sectors. IT is characterized not only by a significant increase in the number of new terms, but also by rapid changes in terminology (Jaleniauskiene & Čičelytė 2011, p. 120), because information technology is considered one of the most recent. In the modern context, the rapid development of IT has had a significant impact on linguistic processes, particularly at the lexical level (Burunina & Havrylova 2021, p. 123). The field of artificial intelligence development is no exception; it is advancing rapidly and, as it evolves, is adopting new terminology.

For a more extensive examination of technical terminology, we will consider classification by A. Sydor and R. Nanivsky, in which he identifies the following types of lexical units found in IT texts:

1) General literary terms used in their standard meaning. These are common language units that retain their usual meaning and primarily fulfil grammatical or functional roles within the text.
2) General linguistic units used in a specialized context. These words form part of the common language, but take on a different meaning in specialist information technology texts.
3) Phraseological expressions are fixed or semi-fixed expressions, the meaning of which cannot always be fully predicted on the basis of the individual words they consist of, and which can function as established units in speech.

4) General vocabulary that is not usually considered scientific but is relevant to the topic. These are words or phrases that originate from common language but take on a conceptual meaning in the context of information technology.
5) Specialized terminology consists of precise terms that denote specific concepts in the field of information technology and form the basis of its professional vocabulary.

In our study of the scientific terminology in the book *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*, we examine four categories of specialized terms.

Let us consider the examples of general linguistic units used in a specialized context from the text:

(1-s) *"For years the field had been working to merge the first two — language and vision — so a single* ***model*** *would be capable of relating words to visual information."(Hao 2025, p. 226) —* **(1-t)** *Упродовж років дослідники намагалися об'єднати мову й зір, щоб одна* ***модель*** *могла співвідносити слова з візуальною інформацією*.

In technical texts the word "model" takes on the meaning of a mathematical model used by artificial intelligence systems for data processing, prediction, and task execution; it is typically created using machine learning or statistical methods (Glossarytech).

(2-s) *"For DALL-E 2, the research team had signed a licensing deal with stock photo platform Shutterstock and done a massive* ***scrape*** *of Twitter to add to its existing collection of 250 million images." (Hao 2026, p. 229) —* **(2-t)** *У випадку DALL-E 2 дослідницька команда уклала ліцензійну угоду зі стоковою платформою Shutterstock і здійснила масштабний* ***збір даних*** *з Twitter, додавши ці дані до вже наявної колекції з 250 мільйонів зображень*.

In general usage, the term "scrape" refers to the process of removing or collecting something from a surface. However, in the information technology field, it takes on a more specific meaning, referring to the automated extraction of large amounts of online data (Cambridge Dictionary). Thus, although the word itself remains part of the general lexicon, its interpretation depends on the context and has a specific functional meaning.

Here are the examples of phraseological expressions:

(3-s) *"With roadblocks that needed to be punched through in the way of GPT- 4's development****, the stars aligned****." (Hao 2025, p. 235) —* **(3-t)** *І коли на шляху розробки GPT-4 з'явилися бар'єри, які потрібно було буквально проламати,* ***обставини склалися для нього якнайкраще***.

The idiom "the stars aligned" is used in computer science in a metaphorical sense to describe a situation where many factors or conditions coincide, ensuring the successful execution of a complex technological process. Although this expression originates from common language, in this context it refers to the work on GPT-4 to emphasise that various technical, organisational and research challenges were resolved in a timely and favourable manner.

(4-s) *"It* ***rolled out*** *a beta program, inviting one million people around the world to get access to the model with free credits for image generations." (Hao 2025, p. 233) —* **(4-t)** *OpenAI* ***запустила*** *бета-програму, запросивши близько мільйона людей у всьому світі й надавши їм безкоштовні кредити на генерацію зображень*.

The idiomatic expression "rolled out" has been translated as "запустила", which is its functional equivalent. Rather than retaining the metaphorical imagery of the original phrasal verb, we have conveyed its meaning *to launch* or *introduce a product* in a form that is natural and familiar within Ukrainian technical discourse (Cambridge Dictionary).

Here some examples of general vocabulary that is not usually considered scientific but is relevant to the topic from the fragment on the book:

(5-s) *"How fast you think AI will advance and reach major milestones like AGI is your "****AI timeline****."" (Hao 2025, p. 224) —* **(5-t)** *Те, наскільки швидко, на вашу думку, розвиватиметься ШІ й досягне ключових рубежів, зокрема ШІ (AGI), називають «****ШІ-таймлайном****»*.

The text notes that "AI timeline" is not a strictly standardized scientific term, but rather a conceptual idea relating to the development and discourse surrounding predictions about the future

of artificial intelligence.

(6-s) *"The two groups became colloquially known as the* ***Doomers*** *and* ***Boomers****." (Hao 2025, p. 225) —* **(6-t)** *У розмовній мові ці два табори почали називати «****песимісти****» та «****прогресисти****».*

These units are not formalized technical terms, nor are they standard elements of scientific discourse. Instead, the emergence of these terms was caused by the online community, where "doomers" — those with a negative outlook on the development of artificial intelligence — were countered by the emergence of "boomers" (Merriam-Webster).

Specialized terminology from the text:

(7-s) *"This included updated* ***content-moderation filters*** *that wrapped around the model to block abusive images in addition to text as well as a* ***user-behavior-monitoring platform*** *and a so-called* ***ban infrastructure****—systems that automatically suspended user accounts that reached a certain threshold of repeat offenses." (Hao 2025, p. 229) —* **(7-t)** *Це передбачало оновлені фільтри модерації контенту, які працювали поверх моделі й блокували не лише текст, а й небажані зображення, а також систему моніторингу поведінки користувачів і так звану* ***інфраструктуру обмежень*** *— механізми автоматичного призупинення акаунтів після досягнення певного порогу повторних порушень.*

The sentence contains several units of specialized terminology that refer to clearly defined functions of digital systems and artificial intelligence systems. Unlike general vocabulary, these terms are used in professional discourse and relate to specific technical components of the programme. For example, "content-moderation filters" are automated mechanisms for detecting and blocking inappropriate material; "a user-behavior-monitoring-platform" is a system designed to monitor and analyse user actions; and "ban infrastructure" is an integrated set of mechanisms that can be used to apply restrictions to users' accounts.

(8-s) *"****Stable Diffusion****, the popular* ***open-source image generator****, would require only 256* ***Nvidia A100s*** *to train, using a revised technique known as* ***latent diffusion****." (Hao 2025, p. 227) —* **(8-t)** ***Stable Diffusion*** *— популярний* ***генератор зображень з відкритим кодом*** *— потребував для навчання лише 256 графічних* ***процесорів Nvidia A100****, використовуючи вдосконалений підхід, відомий як* ***латентна дифузія****.*

Terms such as "Stable Diffusion and Nvidia A100" are standardized proper nouns denoting specific technological entities — namely, an artificial intelligence model and a graphics processing unit architecture, respectively. Their meaning is unambiguous and universally recognized within the industry, which is a defining characteristic of specialized terminology. "Open-source image generator" and "latent diffusion" denote conceptually distinct concepts, since "latent diffusion" refers to a specific computational method with a defined mechanism and scope of application, whereas "image generator" denotes a specific class of artificial intelligence systems designed to create visual content based on data or prompts.

Thus, based on the examples provided above, the complexity of the artificial intelligence branch lies not only in the presence of highly specialized terms, but also in the fact that many words, which at first glance appear ordinary, take on a precise technical meaning that can easily be misunderstood without sufficient knowledge of the field.

In conclusion, translating texts in the field of artificial intelligence is particularly challenging due to the constant evolution of terminology and its heavy reliance on context. Moreover, the rapid development of this field leads to the constant emergence of new concepts and lexical items, for which there are often no established equivalents in the target language. This requires the translator to rely not only on available resources, but also on analytical skills, the ability to understand context. The statistical analysis conducted in this research revealed the following: general linguistic units used in a specialized context (24%), phraseological expressions (8%), general vocabulary that is not usually considered scientific but is relevant to the topic (16%), specialized terminology (52%).

2.3 Practical analysis of the employed techniques in terms of translating technical terms

This section of our work focuses on the translation of technical terminology and the translation techniques used. It is important to understand the function and impact of a concept before translating it (Newmark 2008, p. 155). For the purposes of this study, the translation techniques proposed by L. Molina and H. A. Albir were employed. Although their classification features 18 techniques, we have used 6 of them in our translation, namely: borrowing, calque, description, discursive creation, modulation, and established equivalent (2002, p. 509). This will ensure terminological accuracy and contextual adaptation.

The first technique is *borrowing*, which involves transferring a lexical unit from the source language to the target language without translation (L. Molina H. & A. Albir 2002, p. 510). This technique is particularly common in technical translation, as specialized terminology is often international in nature and tends to be preserved across languages without significant adaptation.

Here are some examples of *borrowing* in the translation of technical terms:

(9-s) *How could it be weaponized to produce synthetic CSAM or political* ***deepfakes****? (Hao 2025, p. 230) —* **(9-t)** *Чи може модель бути використана для створення синтетичного CSAM або політичних* ***дипфейків****?*

The translation of the word “deepfakes” as “дипфейків” is an example of the use of borrowing, specifically through transliteration. The term is transferred into the target language whilst retaining its original form, which ensures consistency and reflects the term’s widespread use in the modern technological industry. This method is reasonable, as the term has already become established around the world, and there is no equivalent in the target language.

(10-s) *While the first* ***Transformer*** *had been initially designed to work best with text, Google had introduced a new Vision Transformer in 2020, adapting it to images (Hao 2025, p. 226). —* **(10-t)** *Хоча перші* ***трансформери*** *спершу створювалися передусім для роботи з текстом, у 2020 році Google представила Vision Transformer — адаптовану версію цієї архітектури для обробки зображень.*

The term “Transformer” is rendered as “трансформери”, which illustrates the use of borrowing through transliteration. The term is translated into the target language whilst retaining its exact form, which maintains terminological accuracy and consistency within the field of artificial intelligence. This approach is justified by the absence of a generally accepted equivalent in the Ukrainian language and the established use of the borrowed form in a professional context.

Calque is a literal translation of a foreign word or phrase, it can be lexical or structural. Lexical calque involves translating the individual components of a term directly into the target language. Structural calque involves reproducing the grammatical structure of the source language term (L. Molina, H. A. Albir, 2002, p. 510). Now, let us consider the following examples of calque:

(11-s) *What had, on the face of it, been OpenAI’s easiest goal in its 2021 research road map turned out to be one of the hardest: scaling up GPT-3 by 10x with Microsoft’s new eighteen thousand Nvidia A100* ***supercomputer cluster****, in its effort to develop what would become GPT-4 (Hao 2025, p. 233). —* **(11-t)** *Те, що на перший погляд здавалося найпростішим пунктом у дослідницькій дорожній карті OpenAI на 2021 рік, виявилося одним із найскладніших: масштабувати GPT-3 у десять разів, використовуючи новий* ***суперкомп’ютерний кластер*** *Microsoft із вісімнадцяти тисяч графічних процесорів Nvidia A100, у межах розробки того, що згодом стане GPT-4.*

The translation of the term “supercomputer cluster” as “суперкомп’ютерний кластер” is a calque, as both the lexical components and the syntactic structure of the original term are directly reproduced in the target language.

(12-s) *In the same way DALL-E could generate an avocado armchair having only ever seen avocados and armchairs, DALL-E 2 and DALL-E 3 could do the same thing with children and porn for child pornography, a capability known as “****compositional generation****.” (Hao 2025, p. 229) —* **(12-t)** *Аналогічно до того, як DALL-E могла створити «крісло з авокадо», маючи у своєму розпорядженні лише зображення авокадо й крісел, DALL-E 2 та DALL-E 3 були здатні поєднувати окремі візуальні категорії — наприклад, дітей і порнографічний*

*контент — у нові зображення. Така здатність називається «**композиційною генерацією**».*

We have translated the term "compositional generation" as "композиційна генерація", which is a calque, lexically and structurally the term remains the same as in the source language and is transferred into the target language without any changes to its form. This strategy ensures terminological clarity and accuracy, as each element of the compound term corresponds to its Ukrainian equivalent, and the term itself corresponds to established usage in a technical context.

The next technique is *description*, which entails replacing a term or expression in the source language with a descriptive phrase in the target language that conveys its meaning or intended purpose (L. Molina & H. A. Albir 2002, p. 510). Here are some examples:

(13-s) *... as is "**AI takeoff**," the process of AGI improving to the point of superintelligence and thus capable enough to outwit humanity (Hao, 2025, p. 224).* — **(13-t)** *... та «**швидкий перехід штучного інтелекту до надінтелекту**» — процес, у якому ШІ стає достатньо потужним, щоб перехитрити людство.*

The term "AI takeoff" has been translated using a descriptive approach, as we have conveyed its meaning through phrase rather than equivalent term. This has preserved the clarity and purpose of the term, given that there are no direct equivalents in Ukrainian.

(14-s) *After GPT-3, researchers had sought to accumulate as much data as possible, building up the company's reservoir by downloading every new **data dump** and scraping every new online forum they stumbled upon that didn't have clear warnings against doing so (Hao 2025, p. 234).* — **(14-t)** *Після GPT-3 дослідники намагалися зібрати якомога більше даних, розширюючи «резервуар» компанії: завантажували новий **великий масив даних** й аналізували кожен новий онлайн-форум, на який натрапляли, якщо там не було чітких застережень проти такого збору даних.*

The term "data dump" has been translated as "великий масив даних", a descriptive approach that conveys the meaning of this informal and context-dependent term through an explanatory phrase rather than a direct translation.

Regarding the subject of our study, technical translation, certain terms have already come into common usage. It is in such cases that we will apply the *established equivalent* technique, using standardized rendering options found in dictionaries or direct equivalents (L. Molina, H. A. Albir, 2002, p. 510). For example:

(15-s) *To live in San Francisco and work in **tech** is to confront daily the cognitive dissonance between the future and the present, between narrative and reality (Hao 2025, p. 219).* — **(15-t)** *Життя у Сан-Франциско й робота у **сфері технологій** надає можливості щодня стикатися з когнітивним дисонансом між майбутнім і сьогоденням, між ідеєю та реальністю.*

The translation of "tech" as у "сфері технологій" demonstrates the use of an established equivalent, as our rendering is conventional, semantically accurate, and stylistically appropriate in Ukrainian. According to the Cambridge Dictionary, "tech" is a shortened form of the words "technical" or "technology" and refers to matters relating to science, technology and industry. Therefore, the equivalent we have chosen accurately conveys the general meaning of the source term, whilst corresponding to the norms of formal discourse in the target language.

Here is another example of *established equivalent* in our translation:

(16-s) *The share of pornographic images on the internet was so large that removing them shrank the **training dataset** enough to notably degrade the model's performance (Hao 2025, p. 229).* — **(16-t)** *Частка порнографічного контенту в Інтернеті виявилася настільки значною, що його вилучення суттєво зменшувало **тренувальний набір даних** і помітно погіршувало якість роботи моделі.*

The government-approved glossary of artificial intelligence terms provides standard translations: "training data" is translated as "тренувальні дані", whilst the general term "dataset" corresponds to "набір даних" or "датасет" (2024). Therefore, the combined phrase "training dataset" as "тренувальний набір даних" a correct, standardized expression, rather than a context-dependent translation, which confirms its role as a generally accepted equivalent in technical

discourse.

Modulation is a translation technique involving an alteration of perspective, emphasis or conceptual class compared to the source text, which may occur at the lexical or structural level (L. Molina, H. A. Albir, 2002, p. 510).

Let us provide some examples of the *modulation* technique:

(17-s) *In early 2022, OpenAI was ready to test a different product release strategy, this time with its* ***text-to-image work*** *(Hao 2025, p. 225).* — **(17-t)** *На початку 2022 року OpenAI була готова випробувати нову стратегію запуску продукту — цього разу щодо своєї* ***текстово-візуальної моделі****.*

Our use of the term "текстово-візуальної моделі" to refer to "text-to-image work" illustrates *lexical modulation*, as the translation replaces the concept of an ongoing process with its final result, ensuring greater precision and compliance with the context of technical discourse in target language.

(18-s) *DALL-E had spun out of a trend in the broader field of AI research to develop multimodal models—****models*** *that combine at least two different "modalities," such as text, images, sound, or video (Hao 2025, p. 226).* — **(18-t)** *DALL-E з'явилася в руслі ширшого тренду в дослідженнях ШІ — створення мультимодальних моделей, тобто* ***систем****, що поєднують щонайменше дві різні «модальності»: текст, зображення, звук або відео.*

The translation of the word "models" as "систем" reflects a lexical change view, as this concept is presented from a different perspective, we have chosen to move from an abstract term to a more specific meaning.

Discursive creation is a translation technique that involves establishing a temporary equivalence between elements of the source and target texts (L. Molina, H. A. Albir, 2002, p. 510).

(19-s) "***Hardware overhang****," as referenced in OpenAI's 2021 research road map, is another dictionary entry… (Hao, 2025, p. 224).* — ***(19-t) «Надлишок обчислювальних потужностей»*** *(термін, згаданий у дослідницькій дорожній карті OpenAI 2021 року)…*

As there is no exact equivalent for this term in Ukrainian, we have chosen a contextual translation that explains the concept using the phrase "надлишок обчислювальних потужностей". This approach preserves the communicative function of the original term and ensures clarity for the target audience.

Thus, the analysis demonstrates that translating AI terminology requires a flexible, context-dependent approach. Examples show translation requires a variety of strategies to be effective. This approach is intended to achieve precision, clarity, and consistency in the rendering of both specialized terms and context-dependent lexical units. In the course of our research, we examined techniques used to render artificial intelligence terminology, including calque (19%), borrowing (29%), description (10%), discursive creation (1%) modulation (14%), and established equivalent (27%) (Appendix B). Our analysis of the selected examples demonstrates that the choice of translation techniques depends on the term's semantic and structural features, its level of standardization, and its contextual function. The results show that established equivalents and borrowing are most commonly used, primarily because the terminology for AI is still being developed in Ukrainian.

Conclusions

The focus of our translation project is the analysis and translation of artificial intelligence terminology based on the non-fiction book *Empire of AI: Dreams and Nightmares at Sam Altman's OpenAI* by Karen Hao. We analysed and translated selected excerpts from the non-fiction text, focusing on the peculiarities of rendering AI-related terms into Ukrainian. Special attention was paid to preserving semantic accuracy and the stylistic features of the source text, including its analytical and narrative character. According to the research objectives:

1. The types and structural characteristics of the lexical units found in the text were analysed and then classified according to A. Sydor and R. Nanivsky. The results are as follows: general linguistic units used in a specialized context (24%), phraseological expressions (8%), general vocabulary relevant to the topic (16%), and specialized terminology (52%) (Appendix A). This classification shows that AI discourse is characterised by specialised terminology and context-dependent units.
2. The main challenges that arise when translating AI terminology were examined. These include the lack of standard Ukrainian equivalents, the rapid emergence of new terms, and the semantic ambiguity of a number of lexical units. These challenges complicate the process and require careful analysis to avoid misinterpretation.
3. The translation strategies and techniques used to render AI terminology into Ukrainian were defined and analysed. Based on the classification proposed by L. Molina and A. Hurtado Albir, the following techniques were identified (Appendix B): calque (19%), borrowing (29%), description (10%), discursive creation (1%) modulation (14%), and established equivalent (27%) (Appendix B). The results show that a combination of strategies should be used depending on the nature of the term and its role in the text.
4. A practical analysis of translation examples was conducted to determine their appropriateness. The results show that translation requires a balance of terminological accuracy and clarity for the target audience, as well as consistency throughout the text. Newly formed terms and expressions in context should be given particular attention.

Prospects for further research may include a more detailed study of how newly emerging artificial intelligence terminology is being adapted into Ukrainian. It is also important to investigate the consistency of AI terminology usage in Ukrainian translations and to identify ways to improve its standardisation.

In conclusion, translating artificial intelligence terminology requires a comprehensive, context-sensitive approach. As demonstrated in this translation project, the effective translation depends not only on conveying the literal meaning but also on preserving the functional and contextual significance of terms.

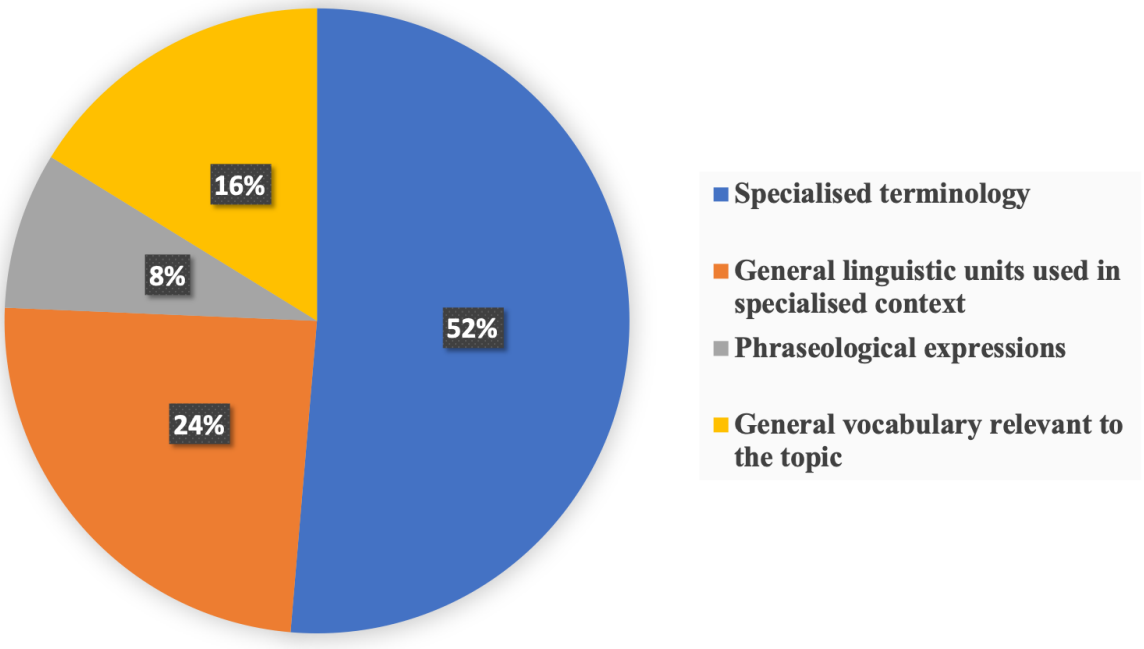
References

1. Burunina, N. V., & Havrylova, I. M. (2021). Translation of scientific and technical texts in the field of IT. *Transcarpathian Philological Studies*, (15), 122–126. https://doi.org/10.32782/tps2663-4880/2021.15.22
2. Byrne, J. (2014). *Scientific and technical translation explained: A nuts and bolts guide for Beginners*. Routledge.
3. Cabré, M. T., & Sager, J. C. (1999). *Terminology: Theory, methods, and applications*. J. Benjamins Pub. Co.
4. Hao, K. (2025). *Empire of AI: Dreams and nightmares in Sam Altman's OpenAI*. Penguin Press.
5. Jaleniauskienė, E., & Čičelytė, V. (2011). Insight into the latest computer and Internet terminology. Studies About Languages, 0(19). https://doi.org/10.5755/j01.sal.0.19.955
6. Merriam-Webster. (n.d.). *Doomer slang meaning*. Merriam-Webster. Retrieved April 7, 2026, from https://www.merriam-webster.com/slang/doomer
7. Molina, L., & Hurtado Albir, A. (2004). Translation techniques revisited: A dynamic and functionalist approach. *Meta*, *47*(4), 498–512. https://doi.org/10.7202/008033ar
8. Newmark, P. (2008). *A textbook of translation*. Pearson Education.
9. Pavel, S., & Nolet, D. (2001). *Handbook of terminology. Terminology and Standardization,* Translation Bureau.
10. *Cambridge Dictionary | English Dictionary, Translations & Thesaurus*. (2026). Retrieved March 21, 2026, from https://dictionary.cambridge.org/
11. *Tech recruiters' search tool for exploring IT terms*. Glossarytech. Retrieved April 3, 2026, from https://glossarytech.com/search/results?phrase=model.html
12. Сидор, А. Р., Нанівський Р. С. (2019) Урахування лексичних особливостей сфери інформаційних технологій під час перекладу з англійської мови українською. Закарпатські філологічні студії. Ужгород. Retrieved March 13, 2026, from https://dspace.uzhnu.edu.ua/jspui/handle/lib/27798
13. Синдега, Р. Є., & Іващишин, О. М. (2009). Структурні особливості функціонування термінів в англомовних текстах з проблем комп'ютерних наук та інформаційних технологій. Наукові записки [Національного університету Острозька академія]. Сер.: Філологічна, (11), 351–358.
14. СЛОВНИК ТЕРМІНІВ У СФЕРІ ШТУЧНОГО ІНТЕЛЕКТУ. Retrieved March 17, 2026, from https://storage.thedigital.gov.ua/files/2/72/389a01ab0cc82040dfe172f94d1af720.pdf

Appendices

Appendix A

Classification of technical terminology by A. Sydor and R. Nanivsky



Appendix B

Techniques of translating technical terms (based on the classification by L. Molina and H. A. Albir)