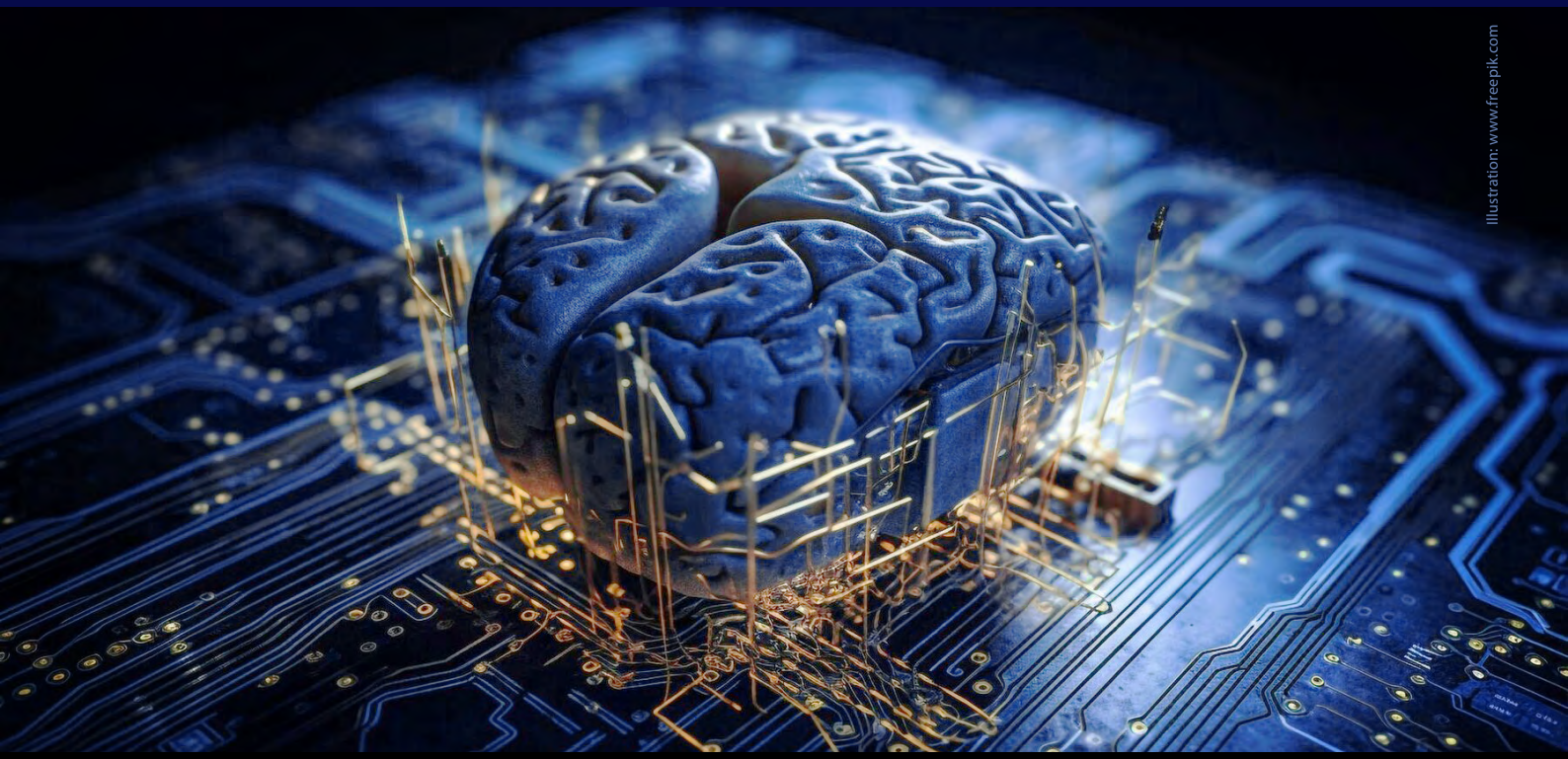

Illustration: www.freepik.com

AI SAFETY SUMMIT

ГЛОБАЛЬНИЙ САМІТ З БЕЗПЕКИ ШТУЧНОГО ІНТЕЛЕКТУ

1 листопада в Блетчлі-парку (Велика Британія) почав роботу перший у світі глобальний саміт з безпеки штучного інтелекту (ШІ). Його мета — вивчити ризики, пов'язані з технологією, що швидко розвивається, і дати поштовх міжнародному діалогу щодо її регулювання.

У саміті взяли участь світові політичні лідери, серед яких: віцепрезидент США Камала Гарріс, президент Європейської комісії Урсула фон дер Ляйєн, віцеміністр технологій Китаю У Чжаохуей і генеральний секретар ООН Антоніу Гутерреш, а також представники технологічних компаній, учені та некомерційні організації. Серед учасників – керівники найвідоміших у світі компаній, що займаються штучним інтелектом, у тому числі генеральний директор Google DeepMind Деміс Гассабіс і Сем Альтман, який заснував компанію OpenAI — розроблювача ChatGPT, представники компаній Alibaba та Tencent і звичайно ж мільярдер Ілон Маск.

Фото: https://theconversation.com/

Мета саміту — почати глобальну розмову про майбутнє регулювання ШІ

Незважаючи на те, що минув майже рік з того часу, як OpenAI представила публіці ChatGPT — чат-бот зі штучним інтелектом, поки не існує всеосяжних глобальних правил, що стосуються безпеки ШІ. Уряди деяких країн звернули на це увагу та почали розробляти свої власні правила. Наприклад, Європейський Союз видав перший збірник законів, що регулюють використання ШІ.

У рамках саміту відбулася серія круглих столів, на яких обговорювалися загрози, пов'язані з майбутнім розвитком технологій ШІ. Наприклад, використання ШІ хакерами для DDos-атак або терористами для створення біологічної зброї та завдання збитків світу.

Експерти та регулювальні органи поки розходяться в думках щодо того, як розставляти пріоритети стосовно цих загроз: у довгоочікуваному Законі ЄС про штучний інтелект говориться про можливі порушення прав людини, таких як конфіденційність даних і стеження, а не про так звані екзистенційні ризики, які домінують у більшій частині порядку денного саміту.

Прем'єр-міністр Великої Британії Ріші Сунак
бачить Велику Британію світовим лідером у сфері безпеки штучного інтелекту. Він планує створити глобальну консультативну раду, яка регулюватиме ШІ за зразком Міжурядової групи експертів зі зміни клімату (IPCC). Передбачається, що Британія буде відігравати роль посередника між трьома блоками великих держав світу — США, ЄС і Китаєм, і саме перший саміт з безпеки ШІ закладе основу для майбутнього міжнародного діалогу.

Міністр торгівлі США Джина Раймондо
заявила, що Сполучені Штати створять Інститут безпеки штучного інтелекту для оцінки відомих і можливих ризиків, пов'язаних з так званими «прикордонними» моделями штучного інтелекту. Вона сказала, що США установлять офіційне партнерство з Інститутом безпеки Об'єднаного Королівства. Дана ініціатива буде здійснюватися Національним інститутом стандартів і технологій (NIST) і очолить зусилля уряду США із забезпечення безпеки ШІ, особливо його провідних моделей.

Президент США Джо Байден
напередодні заходу підписав указ про штучний інтелект, що вимагає від розроблювачів систем штучного інтелекту, які можуть являти загрозу для національної безпеки, економіки, суспільної охорони здоров'я та розвитку США, ділитися результатами випробувань щодо безпеки з урядом США, так само, як це прийняте в оборонному виробництві. Указ також пропонує агентствам установити стандарти для цих випробувань і усунути пов'язані з ними хімічні, біологічні, радіологічні, ядерні ризики та ризики кібербезпеки.

Ілон Маск
заявив, що перший саміт з безпеки штучного інтелекту у Великій Британії має обговорити створення органа незалежної експертизи, який міг би контролювати компанії, що розробляють штучний інтелект. «Насправді необхідно, щоб існував незалежний експерт, який міг би спостерігати за тим, що роблять провідні компанії у сфері штучного інтелекту, і, принаймні, бити тривогу, якщо з'являться проблеми, — сказав підприємець-мільярдер журналістам у Блетчлі-парку. — Я не знаю, які правила обов'язково мають бути введені, але перш ніж приступати до нагляду, потрібно розуміти, як це зробити», — сказав Маск.

За результатами роботи саміту була опублікована декларація

[https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023], **підписана представниками 28 країн, включно зі США та Китаєм, а також Європейським Союзом, у якій викладено програму, спрямовану на виявлення ризиків, пов'язаних із ШІ, що викликають загальну стурбованість, формування наукової термінології для однозначного їх розуміння та розробку політики протидії їм. Програма одержала назву «Декларація Блетчлі» та є по суті відправною точкою глобального обговорення передбачуваної небезпеки ШІ. Декларація містить ключові завдання для розробки спільної угоди про відповідальність стосовно ризиків, а також подальшої міжнародної співпраці у сфері безпеки штучного інтелекту.**

На найближчі 12 місяців визначено такі пріоритети міжнародної співпраці у сфері провідного ШІ:

• виробити загальне розуміння можливостей штучного інтелекту та ризиків, які він представляє для глобальної безпеки та благополуччя людей;

• розробити скоординований підхід до досліджень безпеки та оцінок моделей провідних систем штучного інтелекту, включно зі способами їх застосування;

• розвивати міжнародну співпрацю та партнерство, спрямоване на забезпечення того, щоб переваги ШІ були доступні для того, аби звузити глобальну нерівність, а не збільшити її;

• ці пріоритети можуть бути реалізовані на багатосторонніх форумах. Усі повинні працювати разом, щоб забезпечити взаємодоповнюваність і цілеспрямованість різних ініціатив.

Department for Science, Innovation and Technology

Основні заяви учасників саміту

• Роль наукової співпраці полягає в розробці тестів, які продемонструють безпеку ШІ політикам.

• Нам слід використовувати безліч методологій, включно із соціальними науками, оскільки це соціотехнічне завдання. Нам необхідно визначити найбільш важливі питання та зосередитися на них; швидкість має вирішальне значення.

• Системи штучного інтелекту за своєю суттю є міжнародними: створені в одній країні, вони можуть бути легко та швидко розгорнуті в іншій.

• Баланс ризиків і можливостей — складне завдання, враховуючи швидкі темпи розвитку ШІ. Для його успішного досягнення потрібно, щоб регулювання й інновації йшли пліч-о-пліч. Вони не перебувають на протилежних кінцях спектра, і регулювання може стимулювати інновації, у тому числі за допомогою законів про безпеку розробок.

• Підвищена увага до кібербезпеки, включно з принципами безпечного проєктування, є основним критерієм для всіх провідних розроблювачів ШІ.

• Відомий передовий ШІ створює соціальні ризики, які являють собою екзистенційну загрозу демократії, правам людини, цивільним правам, справедливості та рівності (наприклад, економічним можливостям, охороні здоров'я та розвитку).

• Не можна упускати можливість використання ШІ для вирішення глобальних проблем, включно зі зміцненням демократії, подоланням кліматичної кризи й усуненням соціальних упереджень.

• Навіть коли здається, що системи ШІ демонструють високі когнітивні здатності, ми не можемо бути впевнені в тому, що вони будуть поводитися так само або ухвалювати ті ж рішення, що й люди. Наприклад, майбутні системи без належного контролю можуть діяти так, як їх розроблювачі не очікували або не планували. Зараз ми можемо розпочати конкретні дії, щоб запобігти подібним сценаріям. Є рішення, які не варто передавати системі штучного інтелекту, щоб ми, як суспільство, могли уникати надмірної залежності. Нам також необхідно ретельно тестувати моделі в безпечних середовищах і прогнозувати ситуації, які можуть відбутися в результаті втрати контролю.

• Нинішні можливості провідних систем штучного інтелекту вже серйозно перевершують ті можливості, які багато фахівців пророкували всього кілька років тому.

• У міру збільшення інвестицій досить імовірно, що ми й далі будемо дивуватися тому, на що здатні системи ШІ, причому не обов'язково це буде передбачено або задумано їхніми творцями. Ці моделі також зможуть підключатися до інших систем і розширювати їхні можливості — кількість можливих перестановок важко вгадати, як, відповідно, і потенційні результати.

• Провідні можливості штучного інтелекту, імовірно, принесуть величезну користь, вирішуючи існуючі сьогодні завдання у сфері охорони здоров'я, освіти, навколишнього середовища, науки та в інших сферах. Але ті ж самі властивості систем штучного інтелекту, які створюють ці переваги, створюють і значні ризики.

• Нові передові моделі штучного інтелекту повинні розроблятися та ретельно тестуватися в безпечних умовах. Обіцянка потенційних переваг не повинна бути приводом для пропуску або поспішного проведення випробувань на безпеку або інших оцінок.

• Хоча моделі відкритого доступу мають деякі переваги, такі як прозорість і можливість проведення досліджень, відмовитися від подібної моделі з небезпечними можливостями після її випуску неможливо. Це викликає особливу стурбованість у зв'язку з потенціалом моделей відкритого доступу, які можуть призвести до неправильного використання ШІ, хоча необхідна відкрита дискусія, щоби збалансувати ризики та вигоди.

• Обмін інструментами оцінки — це добре, але це не означає, що ми звільнимося від ризиків. Нам необхідно постійно стежити за ризиками, що виникають.

• Новітні провідні системи штучного інтелекту (GPT4 і еквіваленти) можуть полегшувати навіть не дуже підготовленим зловмисникам проведення кібератак, розробку біологічної або хімічної зброї.

• Передові системи штучного інтелекту, швидше за все, стануть більш функціональними і точними, а також більш розповсюдженими та доступними для зловмисників, тому ці ризики будуть зростати.

• Ризики, які ці системи ШІ представляють для суспільства, значні. Дуже важливо їх досліджувати та знайти способи гарантувати, що нинішні та майбутні моделі не будуть доступні зловмисникам.

Докладніше https://www.aisafetysummit.gov.uk/policy-updates/#company-policies