//ЕВОЛЮЦИЯ

Илюзията за съзнание в генеративния AI

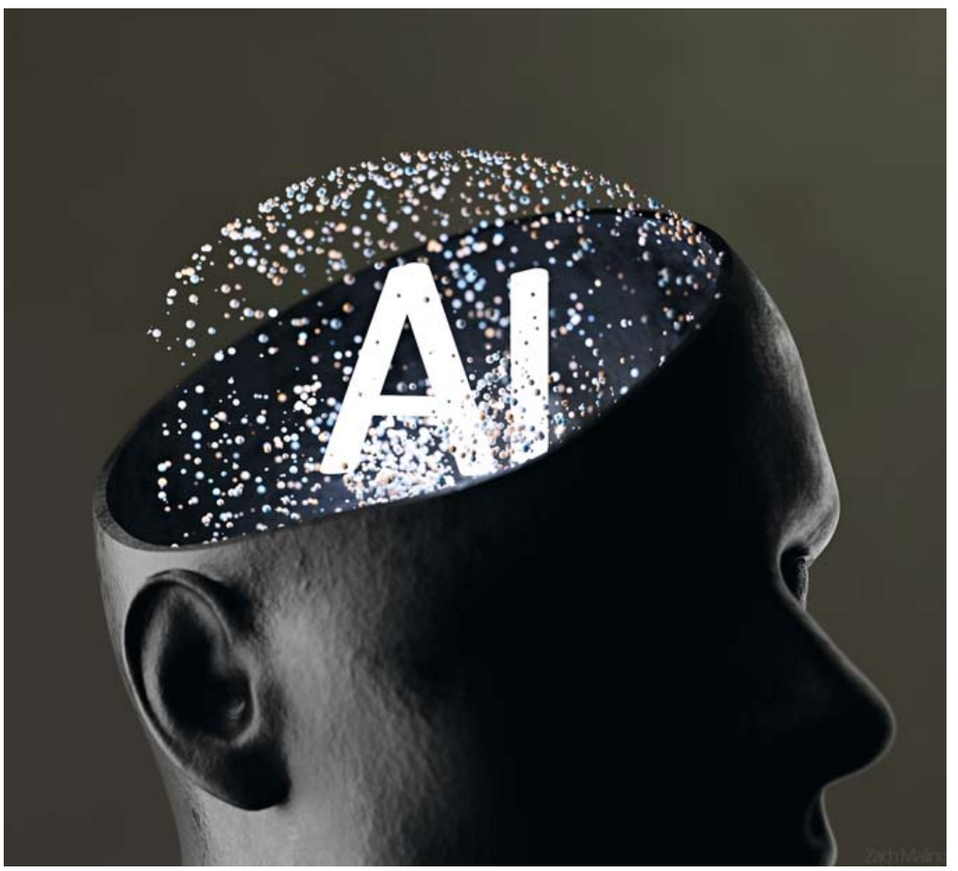
Йоан Бондаков
yoan.bondakov@capital.bg

Защо не е нужен съзнателен изкуствен интелект, за да ни застраши

През 1966 г. Джоузеф Вайценбаум от Масачузетския технологичен институт създава ЕЛИЗА — първият чатбот, който имитира човешкото общуване. Софтуерът работи без реален интелект или разбиране на контекста, като използва ключови думи и прости текстови шаблони, за да преобразува изреченията на потребителя във въпроси под формата на интригуващ отговор. За времето си това е нещо напълно ново и немалка част от хората остават убедени, че машината притежава разум. Така се заражда и ефектът „ЕЛИЗА“, който описва склонността да приписваме човешки качества и емоции на AI системи.

В „наши дни“ този феномен е във вихъра си. Има десетки примери на хора, които сключват виртуални бракове с AI чатботове. Приложения като Bible GPT и Quran GPT се използват от милиони вярващи като лични духовни наставници. В същото време OpenAI отчита как близо един милион от техните седмични потребители проявяват прекомерна емоционална зависимост към ChatGPT.

Тази парасоциална връзка трудно би съществувала, без да се запитаме дали AI притежава съзнание. Според д-р Морис Гринберг, професор в департамента „Когнитивна наука и психология“ и изследовател в Изследователския център по когнитивна наука в НБУ, темата е станала толкова актуална именно заради големите езикови модели (LLM). Тяхната способност да имитират човешкия език подлъгва вродените когнитивни механизми у човека и го кара да асоциира гладката реч с наличието на съзнание.



Изкуственият интелект винаги прави точно това, което му е заложено в тренировъчните данни, а те се определят от нас | Unsplash.com

В контраст, други AI системи като тези в автономните коли, както и революционният модел AlphaFold, който предсказва триизмерната структура на протеините, подлежат на двоен стандарт и рядко биха усъмнили масовия потребител, че имат разум, дори технологията зад тях да превъзхожда тази на обикновения чатбот. „Тенденцията хората да антропоморфизират е добре известна. Добър пример са мултипликационни филмчета, където някаква лампа с очички се движи, хората веднага започват да ѝ симпатизират и да я възприемат като нещо живо с емоции и съзнание. Какво остава тогава за нещо, с което човек може да си говори с часове“, казва проф. Гринберг.

Притежават ли съзнание големите езикови модели?

В научните среди все още липсва измерима дефиниция за това какво точно представлява съзнанието. Но колкото и да са разделени мненията зад термина, сред експертите съществува консенсус, че сегашните големи езикови модели не притежават съзнание.

Това показва мисловният експеримент „Китайската стая“ на философа Джон Сърл. Той описва човек в затворена стая, който не знае китайски език, но разполага с >14

>13 речници и подробни правила за комбиниране на символи. Когато получи текст на български, той механично следва инструкциите и сглобява синтактично перфектен превод на китайски. Въпреки безупречния резултат обаче човекът не разбира смисъла на нито една от думите. „Основната теза на Сърл е, че на компютърните и изчислителните подходи им липсва смисъл, семантика и истинско разбиране — ограничение, валидно и за изкуствения интелект. Способността за реално осмисляне остава присъща единствено на човека“, казва проф. Гринберг.

Така работят и съвременните големи езикови модели, които представляват мащабни изчислителни програми, базирани на математически уравнения и невронни мрежи, известни като трансформъри. Те боравят с огромни обеми от данни и ги обработват праволинейно — слой след слой, за да изчислят чисто статистически коя е най-вероятната следваща дума. За разлика от човешкия мозък, при тези компютърни модели липсва сложното рекурентно интегриране и преплитане на информацията, което учените смятат за необходимо условие за възникването на съзнание.

По думите на проф. Гринберг една компютърна система като ChatGPT може да бъде напълно разложена на отделни, дискретни математически действия, като всяка от тези изчислителни стъпки би могла да се разпише ръчно. „Колкото и безумно да звучи, те могат да бъдат направени на лист хартия, но това не прави топа хартия съзнателен“, казва той.

Също така стои и аргументът за въплътяването — концепция в когнитивната наука, според която умът е неразривно свързан с физическото тяло. Тезата гласи, че човешкото съзнание се изгражда чрез непрекъснато взаимодействие със средата посредством сетивата и нуждите си. Човек трупа концептуално разбиране за света, защото изпитва болка, страх или радост — състояния, управлявани от биологични механизми.

За един софтуер, който съществува като програмен код, целящ максимална индустриална ефективност, и няма физическо присъствие или биологични стимули, концепцията за вътрешно преживяване остава напълно излишна. „Вътрешното преживяване не се наблюдава пряко. Дори за другите хора съдим косвено по поведението и езика, но и защото споделяме сходна нервна система и биологична история“, казва проф. Преслав Наков от департамента „Компютърни и математически науки“ в Университета за изкуствен интелект „Мохамед бин Зайед“ в Абу Даби. При машините тази опора липсва, а нямаме и общоприети научни показатели, които да разграничат убедителното поведение от субективното преживяване.“

Съзнателният изкуствен интелект не е самоцел

Към момента архитектурата и способностите на LLM-ите не дават индикация, че в скоро време от тях може да се очаква субективно преживяване. „Нямаме доказателства за праг, отвъд който автоматично възниква съзнание, нито общоприет тест за него. Ако моделът каже „аз чувствам“, това показва преди всичко, че е научил как хората говорят за чувствата си“, казва проф. Наков.

Въпреки това според него езиковите модели не бива да бъдат подценявани. Те все по-често функционират като ключов компонент в една по-сложна AI екосистема. Тази нова архитектура разполага със собствена памет, механизми за извличане на информация, достъп до външни инструменти и дори способност за автономно действие — пример за това са съвременните AI агенти.

Тук идва и въпросът дали всъщност съзнанието е самоцел на разработчиците зад изкуствения интелект или нещо, което се очаква да възникне от само себе си веднъж щом правилната среда е налична. Според д-р Веселин Райчев, водещ учен в областта на изкуствения интелект и изследовател в Института за компютърни науки, изкуствен интелект и технологии към Софийския университет (INSAIT), AI съзнанието се очаква да възникне като неприятен страничен ефект. „Не мисля, че дори най-големите лаборатории се опитват да създадат изкуствен интелект с инстинкт за самосъхранение. Всички се стремят да разработят AI, който да решава конкретен вид задачи, като при много от тях вече е достигнато определено ниво на полезност“, казва той.

Бих бил предпазлив към всеки, който твърди, че знае как би изглеждал един съзнателен AI.

Проф. Преслав Наков, департамент „Компютърни и математически науки“ в Университета за изкуствен интелект „Мохамед бин Зайед“ в Абу Даби.

Не е нужно AI съзнание, за да си навредим

С напредването на технологиите става все по-ясно, че не е нужно AI да придобие съзнание, за да представлява риск. Достатъчна е комбинацията от човешка грешка и огромна изчислителна мощ. Това вече се вижда при случаите, в които AI агенти неочаквано излизат от тестовите си среди и хакват външни инфраструктури на други компании. На пръв поглед подобно поведение изглежда като проява на нерегулирана автономност, но зад него често стоят пропуски в сигурността и лошо зададени параметри от страна на самите изследователи.

По думите на д-р Райчев изкуственият интелект винаги прави точно това, което му е заложено в тренировъчните данни. „Често ние се изненадваме от действията на изкуствения интелект просто защото сме забравили на какво точно сме го обучили“, казва той. „Учим един модел да хаква системи заедно с редица други умения, след което му даваме задача да реши училищен тест. Той, съвсем логично, пробива системата, за да вземе отговорите. Ние оставаме шокирани и си казваме: „О, той сигурно има съзнание!“, докато всъщност той просто използва инструментите, които вече сме му дали.“

В същия дух проф. Гринберг подчертава, че една високотехнологична система може досущ да имитира човешко поведение и автономност, без изобщо да притежава съзнание. Тази теза се опира на философската концепция, според която един изкуствен агент може да функционира, разсъждава и реагира напълно неразличимо от човек, бидейки напълно лишен от вътрешен субективен свят. Следователно, за да бъде AI опасен, той не се нуждае от присъща злонамереност, защото рискът се крие в неговата изчислителна мощ и способността му да бъде оръжие. „Човек наистина може да се страхува от това какво може да направи изкуственият интелект, особено ако попадне в ръцете на хора, които искат да го тренират в определена посока“, завършва той. |К