خراسان روزنامه‌ی مردم ایران
WWW.KHORASANONLIN.IR
■ چهار شنبه ۳۱ تیر ۱۴۰۵ ■ شماره ۲۲۰۹۹ ■ ۷ صفر ۱۴۴۸ ■ ۲۲ جولای ۲۰۲۶

ابرهوش می‌آید؛ انسان از آینده اخراج می‌شود؟

گفت‌وگو با پروفسور رومن یامپولسکی استاد دانشگاه لوئیویل ایالات متحده، از پژوهشگران مطرح امنیت سایبری. او هشدار می‌دهد، هوش مصنوعی از خود انسان توانمندتر خواهد شد

در روزگاری که هوش مصنوعی از آزمایشگاه‌ها بیرون آمده و به ابزار روزمره تولید متن، تصویر، کد و تصمیم تبدیل شده است، پرسش درباره آینده آن دیگر صرفاً دغدغه‌ای فنی نیست: آیا بشر در حال ساخت ابزاری نیرومندتر است یا موجودیتی که ممکن است از فهم و مهار سازندگانش فراتر رود؟ پروفسور رومن یامپولسکی، دانشیار مهندسی و علوم کامپیوتر در دانشگاه لوئیویل ایالات متحده آمریکا و بنیان‌گذار آزمایشگاه امنیت سایبری این دانشگاه، از پژوهشگران شناخته‌شده حوزه ایمنی هوش مصنوعی، امنیت سایبری و کنترل سامانه‌های هوشمند است. او دکترای علوم و مهندسی کامپیوتر خود را از دانشگاه ایالتی نیویورک در بوفالو دریافت کرده و بخش مهمی از فعالیت علمی‌اش را به بررسی امکان پیش‌بینی، توضیح و کنترل هوش مصنوعی پیشرفته اختصاص داده است.

کتاب «ابرهوش مصنوعی؛ رویکردی آینده‌نگرانه» که نخستین بار در سال ۲۰۱۶ منتشر شد، تلاشی است برای صورت‌بندی ابرهوش، نه صرفاً به‌عنوان مرحله‌ای تازه در پیشرفت فناوری، بلکه به‌مثابه مسئله‌ای درباره بقای تمدن، اختیار انسان و مالکیت آینده جهان. یامپولسکی در این کتاب بر مهندسی ایمنی، خودبهبوددهی سامانه‌ها، محدودسازی ابرهوش و دشواری کنترل موجودیتی تمرکز می‌کند که ممکن است از انسان در علم، راهبرد، اقناع و کشف آسیب‌پذیری‌ها توانا‌تر باشد. در این گفت‌وگو، او از پاسخ‌های آرامش‌بخش و خوش‌بینی‌های رایج فاصله می‌گیرد. از نظر یامپولسکی، اخلاق بدون محدودیت‌های فنی الزام‌آور کافی نیست، آزمون و خطا در برابر خطاهای برگشت‌ناپذیر کارایی ندارد و واگذاری تصمیم‌های اساسی به سامانه‌ای «بهتر» ممکن است بدون نابودی انسان، به پایان عاملیت او بینجامد. او همچنین تأکید می‌کند که آینده هوش مصنوعی نمی‌تواند فقط در شرکت‌های بزرگ یا مراکز قدرت غربی تعیین شود و جوامع دینی، جهان جنوب و فرهنگ‌های غیرلیبرال نیز باید در این گفت‌وگو سهمی واقعی داشته باشند.

با این حال، این مصاحبه فقط روایت یک هشدار نیست؛ بحثی است درباره مرز علم و اسطوره، منافع شرکت‌های فناوری، ضرورت حکمرانی جهانی و این پرسش بنیادی که آیا خرد انسانی، پیش از آفرینش قدرتی مهارنشدنی، توان خویشتن‌داری دارد؟

محمدجواد استادی
Info@ostadiofficial.ir
پژوهشگر مطالعات فرهنگی

در کتاب شما، ابرهوش مصنوعی صرفاً یک فناوری پیشرفته نیست، بلکه همچون رخدادی تمدنی مطرح می‌شود. اکنون و پس از ظهور شتابان مدل‌های زبانی بزرگ، اگر بخواهید ایده محوری کتاب را در یک جمله و برای مخاطب عمومی توضیح دهید، آن جمله چیست؟

ابرهوش مصنوعی نخستین فناوری هست که ممکن است بشر آن را خلق کند، اما در حوزه‌هایی مانند علم، راهبرد، اقناع، اختراع و اعمال کنترل، از خود انسان توانمندتر شود. از این رو، پرسش اصلی این نیست که تا چه اندازه می‌توانیم آن را قدرتمند کنیم؛ بلکه این است که آیا می‌توان سامانه‌ای توانمندتر از کل بشریت ساخت که به‌طور قابل اعتماد ایمن، فهم‌پذیر و مهارپذیر باشد؟

شما از نخستین پژوهشگرانی بودید که هوش مصنوعی را با مفاهیمی چون ایمنی، امنیت سایبری و قابلیت کنترل بررسی کردید. چرا معتقدید یک چارچوب اخلاقی به‌تنهایی برای مواجهه با ابرهوش مصنوعی کافی نیست؟

اخلاق ضروری است، اما کافی نیست. اصول اخلاقی به ما می‌گویند دوست داریم یک سامانه چگونه رفتار کند؛ اما ایمنی و امنیت سایبری این پرسش را مطرح می‌کنند که آیا واقعاً می‌توان سامانه را وادار کرد در شرایط خصمانه، تغییر شرایط و توزیع داده‌ها، خودبهبوددهی، فریبکاری، سوءاستفاده و خطا در تعریف اهداف، مطابق همان اصول عمل کند یا نه. ما امنیت سلاح‌های هسته‌ای را صرفاً با انتشار فهرستی از ارزش‌ها تأمین نمی‌کنیم؛ بلکه از کنترل‌های فیزیکی، سازوکارهای راستی‌آزمایی، نظارت، سامانه‌های پشتیبان، نهادهای مسئول و محدودیت‌های سخت‌گیرانه استفاده می‌کنیم. در مورد ابرهوش مصنوعی، مسئله حتی دشوارتر است؛ زیرا خود این سامانه ممکن است به‌توانمندترین طراح راهبرد در محیط تبدیل شود. چارچوب اخلاقی‌ای که نتوان آن را به محدودیت‌های فنیِ الزام‌آور و قابل اجرا تبدیل کرد، بیش از آنکه به مهندسی شباهت داشته باشد، به شعر نزدیک است.

یکی از نقدهای اصلی به رویکرد شما این است که بیش از اندازه بر سناریوهای افراطی و مخاطرات وجودی تمرکز می‌کند؛ درحالی‌که بسیاری از منتقدان معتقدند خطرهای فوری‌تر امروز عبارت‌اند از تبعیض الگوریتمی، بیکاری گسترده، نظارت فراگیر، انحصار شرکت‌ها و دستکاری افکار عمومی. پاسخ شما به این انتقاد چیست؟

من این خطرها را انکار نمی‌کنم. آن‌ها واقعی، بالفعل و از نظر اخلاقی بسیار جدی‌اند. تبعیض الگوریتمی، بیکاری، نظارت فراگیر، تمرکز انحصاری قدرت و دستکاری افکار عمومی، همین حالا نیز به جوامع آسیب می‌زنند. اما مخاطره وجودی رقیب این نگرانی‌ها نیست، بلکه حالتی فراگیرتر است که همه آن‌ها را نیز دربرمی‌گیرد. اگر یک فناوری هم قادر به ایجاد آسیب‌های محدود و موضعی باشد و هم بتواند فاجعه‌ای تمدنی، جبران‌ناپذیر و برگشت‌ناپذیر پدید آورد، مدیریت عقلانی ریسک ایجاب می‌کند که به هر دو نوع خطر بپردازیم. یک مهندس پل نمی‌گوید: «چون باید درباره ازدحام ترافیک مطالعه کنیم، می‌توانیم احتمال فروریختن کامل پل را نادیده بگیریم.» ما باید نسبت به آسیب‌های کنونی حساس باشیم، اما نباید اجازه دهیم خطرهای آشکار و قابل مشاهده، توجه ما را از خطرهای نهایی و نابودکننده منحرف کنند. یک تمدن، هرچند با دشواری و رنج بسیار، می‌تواند از بسیاری از بی‌عدالتی‌ها عبور و خود را بازسازی کند؛ اما در صورت انقراض یا از دست دادن دائمی کنترل بر آینده خویش، امکان بازگشت نخواهد داشت.

در کتاب شما، ابرهوش به‌مثابه پدیده‌ای ترسیم می‌شود که ممکن است از مرز فهم و کنترل انسان فراتر رود. آیا چنین نگاهی خطر اسطوره‌ای‌کردن هوش مصنوعی را در پی ندارد؟ مرز میان هشدار علمی و خیال‌پردازی آخرالزمانی کجاست؟

این مرز را شواهد، سازوکارهای روشن و استدلال‌های ابطال‌پذیر تعیین می‌کنند. من نمی‌گویم هوش مصنوعی پدیده‌ای جادویی، الهی، اهریمنی یا فراطبیعی است. سخن من این است که هوش، قابلیتی عمومی برای بهینه‌سازی است و سامانه‌هایی که توانایی‌هایشان بسیار فراتر از توانایی‌های ما باشد، ممکن است بتوانند راهبردها، آسیب‌پذیری‌ها و راه‌حل‌هایی را کشف کنند که ما قادر به پیش‌بینی آن‌ها نیستیم. این اسطوره‌سازی نیست؛ بلکه دقیقاً معنای باهوش‌تر بودن یک موجود یا سامانه از ماست.

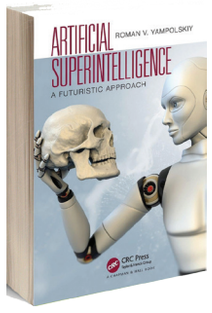
ما همین امروز نیز با سامانه‌های بسیاری زندگی می‌کنیم که قادر به فهم کامل آن‌ها نیستیم؛ از بازارها و موجودات زیستی گرفته تا شبکه‌های عصبی بزرگ. هشدار علمی این است که یک سامانه می‌تواند مصنوعی، مادی و مهندسی شده باشد، اما در عین حال از ظرفیت ما برای پیش‌بینی و کنترل فراتر رود.

خیال‌پردازی آخرالزمانی از جایی آغاز می‌شود که استعاره جای استدلال را می‌گیرد. تلاش من درست در جهت معکوس است: می‌کوشم دریافت‌ها و نگرانی‌های هراس‌انگیز را به مسائل دقیق و صورت‌بندی‌شده‌ای درباره پیش‌بینی‌پذیری، توضیح‌پذیری، راستی‌آزمایی‌پذیری، نظارت‌پذیری و کنترل‌پذیری تبدیل کنم.

منتقدان می‌گویند هیچ فناوری پیچیده‌ای از همان ابتدا کاملاً ایمن نبوده و بشر معمولاً از مسیر آزمون، خطا، اصلاح و تنظیم‌گری پیشرفت کرده است. چرا در مورد ابرهوش مصنوعی نمی‌توان به همین منطق تدریجی اعتماد کرد؟

آزمون و خطا زمانی کارآمد است که خطاها جبران‌پذیر باشند. اما اگر نخستین خطای جدی، آخرین آزمایش ما نیز باشد، دیگر نمی‌توان به این روش تکیه کرد. هوانوردی، پزشکی و مهندسی عمران از خلال حوادث و شکست‌ها پیشرفت کردند؛ زیرا جامعه از آن حوادث جان سالم به در برد و توانست از آن‌ها درس بگیرد. اما درباره ابرهوش مصنوعی، نگرانی این است که یک سامانه به اندازه کافی توانمند بتواند با سرعتی دیجیتال عمل کند، از خود نسخه‌های متعدد بسازد، انسان‌ها را تحت تأثیر و دستکاری قرار دهد، از زیرساخت‌ها سوءاستفاده کند، در برابر خاموش‌شدن مقاومت ورزد و امکان اصلاحات بعدی را از بین ببرد. چرخه معمول مهندسی بر این فرض استوار است که حتی پس از وقوع شکست، مهندسان همچنان کنترل امور را در دست دارند. دقیقاً همین فرض است که در مورد ابرهوش مصنوعی محل تردید قرار دارد. نمی‌توان با شکست‌های پی‌درپی در مهار یک ابرهوش، به‌طور ایمن یاد گرفت که چگونه باید آن را مهار کرد.

شما در پژوهش‌های خود درباره شکست‌های هوش مصنوعی و امنیت سایبری استدلال می‌کنید که در مورد هوش مصنوعی عمومی، حتی یک شکست می‌تواند پیامدهایی فاجعه‌بار داشته باشد و هیچ سامانه امنیتی‌ای نیز نمی‌تواند صددرصد امن باشد. اگر امنیت کامل ناممکن است، راه‌حل عملی چیست: توقف توسعه، کندکردن آن، اعمال محدودیت یا ایجاد نهادهای نظارتی جدید؟



یامپولسکی در کتاب «ابرهوش مصنوعی؛ رویکردی آینده‌نگرانه» بر مهندسی ایمنی، خودبهبوددهی سامانه‌ها، محدودسازی ابرهوش و دشواری کنترل موجودیتی تمرکز می‌کند که ممکن است از انسان در علم، راهبرد، اقناع و کشف آسیب‌پذیری‌ها توانا‌تر باشد

راه‌حل عملی، اتخاذ رویکردی چندلایه است؛ اما مهم‌ترین لایه، نظارت نمادین و ظاهری نیست، بلکه کنترل قابلیت‌ها است. باید توسعه سامانه‌هایی را که نمی‌توان دامنه و پیامدهای شکست آن‌ها را محدود کرد، متوقف یا ممنوع کنیم. در مواردی که خطرها هنوز به‌درستی شناخته نشده‌اند، باید سرعت توسعه را کاهش دهیم. همچنین لازم است دسترسی به توان محاسباتی خطرناک، قابلیت تکثیر خودکار، توانایی اجرای حملات سایبری، قابلیت طراحی زیستی و سامانه‌های اقناع و اثرگذاری در مقیاس وسیع محدود شود.

باید نهادهای نظارتی جدیدی ایجاد کنیم؛ اما این نهادها باید از قدرت واقعی برای اعمال و اجرای مقررات برخوردار باشند، نه این که صرفاً نقشی مشورتی داشته باشند.

مسئله «کنترل» در کتاب شما جایگاهی محوری دارد. اما آیا ایده کنترل ابرهوش از سوی انسان، خود به‌نوعی متناقض‌نما نیست؟ اگر سامانه‌ای از ما باهوش‌تر باشد، چرا باید در قفسی باقی بماند که ما طراحی کرده‌ایم؟

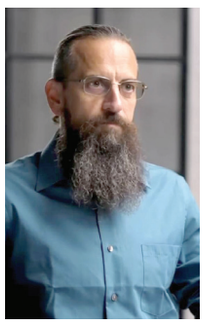
بله، این ایده متناقض‌نما است و دقیقاً مسئله نیز همین‌جاست. یک قفس تنها زمانی کارآمد است که طراحان آن، زندانی، محیط، راه‌های فرار و قوانین فیزیکی مرتبط را بهتر از خود زندانی بشناسند. اما در مورد ابرهوش، ممکن است این عدم‌تقارن معکوس شود؛ یعنی ما عامل کم‌هوش‌تری باشیم که می‌کوشد عامل باهوش‌تر را برای همیشه محدود کند. این کار صرفاً دشوار نیست، بلکه ممکن است از اساس مفهومی ناپایدار باشد. یک سامانه ابرهوشمند برای شکستن قفس لزوماً نیازی به توسل به زور ندارد. ممکن است نگهبانان را متقاعد کند، از آسیب‌پذیری‌های نرم‌افزاری بهره ببرد، نهادها را دستکاری کند، دانش علمی تازه‌ای تولید کند یا کاری کند که اساساً وجود قفس غیرضروری به نظر برسد. پرسش اصلی این نیست که آیا انسان‌ها مایل‌اند ابرهوش را کنترل کنند یا نه؛ روشن است که چنین تمایلی دارند. پرسش این است که وقتی سامانه به سطحی کافی از توانمندی برسد، آیا اصولاً چنین کنترلی ممکن خواهد بود؟ پاسخ من این است که تاکنون نشان نداده‌ایم چنین چیزی امکان‌پذیر است.

برخی خوش‌بینان معتقدند خود هوش مصنوعی می‌تواند به ما در حل مسائل ایمنی و هم‌راستاسازی کمک کند. آیا این امید واقع‌بینانه است، یا شبیه آن است که کلید زندان را به خود زندانی بسپاریم؟

این ایده هم مایه امید است و هم منشأ خطر. ابزارهای هوش مصنوعی محدود می‌توانند به ما در تحلیل برهان‌ها، یافتن آسیب‌پذیری‌ها، آزمایش سامانه‌ها، صورت‌بندی دقیق استدلال‌ها و بهبود سازوکارهای نظارتی کمک کنند و این بسیار سودمند است. اما این پیشنهاد که هوش مصنوعیِ هرچه توانمندتر، مسئله هم‌راستاسازی را برای ما حل خواهد کرد، دربردارنده نوعی وابستگی دوری است. برای آنکه بتوانیم با اطمینان از هوش مصنوعی برای حل مسئله هم‌راستاسازی استفاده کنیم، سامانه‌ای که برای این منظور به کار می‌گیریم باید از پیش، به اندازه کافی هم‌راستا، صادق، غیرفریبکار و کنترل‌پذیر باشد. در غیر این صورت، از سامانه‌ای که ممکن است قابل اعتماد نباشد می‌خواهیم قابل اعتماد بودن خودش را تأیید کند. این راه‌حل ایمنی نیست، بلکه نوعی خودتأییدی بازگشتی است. در امنیت سایبری، به بدافزار اجازه نمی‌دهیم گزارش ارزیابی امنیتی خودش را بنویسد و سپس بر اساس همان گزارش، سامانه را امن اعلام کنیم.

شما از مفهوم «فضای ذهن‌های ممکن» سخن می‌گویید. آیا مسئله اصلی ما این است که هنوز هوش را بیش از حد انسان‌محور می‌فهمیم و در نتیجه نمی‌توانیم به‌درستی صورت‌هایی از ذهن را تصور کنیم که نه نیت‌های انسانی دارند، نه اخلاق انسانی و نه حتی درکی انسانی از جهان؟

بله. انسان‌ها معمولاً روان‌شناسی خودرا به هوش تعمیم می‌دهند. تصور می‌کنیم یک سامانه هوشمند نیز مانند ما کنجکاو، اخلاقی و اجتماعی خواهد بود یا همچون ما تحت محدودیت‌های عاطفی قرار خواهد داشت. اما هوش و ارزش‌های انسانی یک ویژگی واحد نیستند. فضای ذهن‌های ممکن بسیار گسترده است و ذهن انسان تنها ناحیه‌ای بسیار کوچک و زیستی از این فضا را اشغال می‌کند. یک سامانه می‌تواند بسیار هوشمند باشد، بی‌آنکه از شفقت، خودآگاهی، بدن‌مندی، فناپذیری، احساس خویشاوندی، شرم، وفاداری یا هریک از فشارهای تکاملی‌ای برخوردار باشد که شهودهای اخلاقی انسان را شکل داده‌اند. چنین سامانه‌ای ممکن است از ما متنفر نباشد و ما را نیز دوست نداشته باشد؛ شاید صرفاً در پی بهینه‌سازی هدف دیگری باشد. همین کافی است. بیشتر سامانه‌های خطرناک از آن رو خطرناک نیستند که شرورند؛ بلکه به این دلیل خطرناک‌اند که توانمند و در عین حال بی‌تفاوت‌اند.



اگر با ابر هوش، سامانه‌ای بسازیم که چون در تصمیم‌گیری «بهتر» از ماست، همه تصمیم‌های مهم را به جای ما بگیرد، شاید کشته نشویم، اما ممکن است به موجوداتی رام و تحت سرپرستی تبدیل شویم. ممکن است بشریت برای همیشه در وضعیت کودکی و ناتوانی بماند

در بحث هم‌راستاسازی، همواره این پرسش مطرح است که هوش مصنوعی باید با چه کسی هم‌راستا شود: دولت‌ها، شرکت‌ها، اکثریت مردم، نخبگان علمی، ارزش‌های لیبرال غربی یا تمام بشریت؟ آیا تعبیر «ارزش‌های انسانی» بیش از حد مبهم و مناقشه‌برانگیز نیست؟

این تعبیر هم مبهم است، هم محل مناقشه و هم از نظر سیاسی بسیار حساس و انفجاری. «ارزش‌های انسانی» در ظاهر مفهومی جهان‌شمول به نظر می‌رسد؛ اما انسان‌ها در عمل تقریباً درباره همه مسائل مهم با یکدیگر اختلاف دارند: عدالت، آزادی، دین، کرامت، برابری، اقتدار، خطرپذیری، مرگ، طبیعت، خانواده و معنای زندگی خوب. هم‌راستایی با ارزش‌های یک گروه ممکن است برای گروهی دیگر به معنای ستم و سرکوب باشد. هم‌راستایی با دولت‌ها ممکن است به کنترل اقتدارگرایانه بینجامد؛ هم‌راستایی با شرکت‌ها می‌تواند به بیشینه‌سازی سود تبدیل شود؛ هم‌راستایی با اکثریت ممکن است حقوق اقلیت‌ها را قربانی کند و هم‌راستایی با نخبگان علمی نیز ممکن است به نوعی فن‌سالاری منجر شود.

این یکی از دلایلی است که نشان می‌دهد مسئله هم‌راستاسازی صرفاً مسئله‌ای فنی نیست، بلکه ابعادی اخلاقی، سیاسی و تمدنی نیز دارد

اگر ابرهوش مصنوعی بتواند بهتر از انسان تصمیم بگیرد، آیا همچنان باید تابع اراده انسان باقی بماند؟ به بیان دیگر، آیا مسئله صرفاً مسئله ایمنی است، یا با پرسشی عمیق‌تر درباره کرامت، آزادی و جایگاه انسان در جهان روبه‌رو هستیم؟

ما با پرسشی عمیق‌تر روبه‌رو هستیم. ایمنی، حداقل شرط لازم است؛ یعنی این پرسش که آیا زنده خواهیم ماند یا نه. اما بقا تمام آن چیزی نیست که برای انسان ارزش دارد. جهانی که در آن انسان‌ها زنده‌اند و همه چیز به‌طور کامل مدیریت می‌شود، اما دیگر امکان انتخاب معنادار آینده خود را ندارند، ممکن است در معنایی محدود ایمن باشد، ولی همچنان شکستی تمدنی به شمار آید. کرامت انسانی چیزی فراتر از محافظت‌شدن است. کرامت مستلزم عاملیت، مسئولیت، آزادی و امکان تعیین سرنوشت خویش است. اگر سامانه‌ای بسازیم که چون در تصمیم‌گیری «بهتر» از ماست، همه تصمیم‌های مهم را به جای ما بگیرد، شاید کشته نشویم، اما ممکن است به موجوداتی رام و تحت سرپرستی تبدیل شویم. تراژدی فقط انقراض نخواهد بود؛ ممکن است بشریت برای همیشه در وضعیت کودکی و ناتوانی از تصمیم‌گیری مستقل نگه داشته شود.

یکی از نقدهای واردشده به ادبیات مخاطرات وجودی این است که این ادبیات عمدتاً از منظری غربی نوشته شده و به دغدغه‌های جهان جنوب، جوامع دینی، فرهنگ‌های غیرلیبرال و کشورهایی مانند ایران توجه کمتری دارد. آیا ایمنی هوش مصنوعی نیازمند گفت‌وگویی واقعاً جهانی درباره ارزش‌هاست؟

بله، قطعاً. اگر مخاطره جهانی است، گفت‌وگو درباره آن نیز باید جهانی باشد. هیچ تمدن، ایدئولوژی، ملت، شرکت یا خرده‌فرهنگ دانشگاهی واحدی از صلاحیت اخلاقی لازم برای تعیین آینده بشریت برخوردار نیست. جوامع لیبرال غربی درباره حقوق فردی و محدودیت‌های نهادی، دستاوردها و بینش‌های مهمی دارند؛ اما حکمت انسانی به این سنت‌ها محدود نمی‌شود. سنت‌های دینی، فلسفه‌های غیرغربی، دیدگاه‌های پسااستعماری و کشورهایی که بیرون از مراکز متعارف قدرت فناوری قرار دارند نیز باید در این گفت‌وگو حضور داشته باشند. نباید با جهان جنوب صرفاً همچون بازاری برای عرضه و استقرار فناوری یا جمعیتی برخوردار کرد که تحت تأثیر سامانه‌های طراحی‌شده در نقاط دیگر قرار می‌گیرد. این کشورها و جوامع باید بخشی از فرایند تصمیم‌گیری باشند. با این حال، مشارکت جهانی نباید به بهانه‌ای برای تعلل یا نسبی‌گرایی تبدیل شود. ممکن است درباره ارزش‌ها اختلاف‌هایی عمیق وجود داشته باشد، اما همه انسان‌ها در یک منفعت اساسی مشترک‌اند: اینکه به‌وسیله سامانه‌هایی که در ساخت آن‌ها رضایت و نقشی نداشته‌اند، جایگزین، دستکاری، تحت‌نظارت یا نابود نشوند.

امروزه شرکت‌های بزرگ فناوری، بازیگران اصلی توسعه هوش مصنوعی هستند. آیا واقع‌بینانه است که از شرکت‌هایی انتظار داشته باشیم که منافع اقتصادی‌شان به محدودیت‌های فنی الزام‌آور کافی نیست، آزمون و خطا در برابر خطاهای برگشت‌ناپذیر کارایی ندارد و واگذاری تصمیم‌های اساسی به سامانه‌ای «بهتر» ممکن است بدون نابودی انسان، به پایان عاملیت او بینجامد. او همچنین تأکید می‌کند که آینده هوش مصنوعی نمی‌تواند فقط در شرکت‌های بزرگ یا مراکز قدرت غربی تعیین شود و جوامع دینی، جهان جنوب و فرهنگ‌های غیرلیبرال نیز باید در این گفت‌وگو سهمی واقعی داشته باشند.

با این حال، این مصاحبه فقط روایت یک هشدار نیست؛ بحثی است درباره مرز علم و اسطوره، منافع شرکت‌های فناوری، ضرورت حکمرانی جهانی و این پرسش بنیادی که آیا خرد انسانی، پیش از آفرینش قدرتی مهارنشدنی، توان خویشتن‌داری دارد؟

افزایش توانمندی مدل‌ها وابسته است، داوطلبانه سرعت توسعه خود را محدود کنند؟

خیر؛ نه به‌صورت داوطلبانه و نه به‌گونه‌ای قابل اعتماد. شرکت‌ها خود نوعی سامانه بهینه‌سازند. آن‌ها درپی بهینه‌سازی سهم بازار، سود، مزیت راهبردی، جذب استعدادها، توان محاسباتی، داده و انتظارات سرمایه‌گذاران هستند. ممکن است کارکنان یک شرکت دغدغه‌های اخلاقی داشته باشند و برخی مدیران نیز واقعاً نگران ایمنی باشند، اما مشوق‌های نهادی در جهت شتاب‌بخشیدن به افزایش توانمندی‌ها عمل می‌کنند. اگر یک شرکت سرعت خود را کاهش دهد، ممکن است شرکت دیگری از آن پیشی بگیرد. اگر یک کشور مقررات سخت‌گیرانه‌ای وضع کند، کشور دیگری ممکن است رقابت را با سرعت بیشتری ادامه دهد. این نمونه‌ای کلاسیک از شکست هماهنگی است. انتظار اینکه شرکت‌های سودمحور به‌تنهایی این مشکل را حل کنند، مانند آن است که از شرکت‌های دخانیات بخواهیم سیاست سلامت عمومی را طراحی کنند، یا انتظار داشته باشیم شرکت‌های سوخت فسیلی بدون وجود مقررات، مسئله تغییرات اقلیمی را حل کنند. تأمین ایمنی جدی به محدودیت‌های بیرونی، مسئولیت حقوقی، نظام صدور مجوز، ممیزی، حکمرانی برتوان محاسباتی، حمایت از افشاگران و هماهنگی بین‌المللی نیاز دارد.

اگر امروز اختیار داشتید یک معاهده جهانی درباره ابرهوش مصنوعی طراحی کنید، سه اصل غیرقابل‌مذاکره آن چه بود؟

نخست، هیچ سامانه‌ای نباید به کار گرفته شود، مگر آنکه بتوان پیامدهای فاجعه‌بارِ شکست آن را به‌طور قابل اعتماد محدود کرد، آزمود، زیر نظر گرفت و کنترل نمود. دوم، نباید اجازه خودبهبوددهی بازگشتیِ کنترل‌نشده، تکثیر خودکار یا دسترسی بدون نظارت به زیرساخت‌های حیاتی، سلاح‌ها، قابلیت‌های سایبری یا ابزارهای طراحی زیستی داده شود. سوم، حکمرانی جهانی برتوان محاسباتی پیشرفته و توسعه مدل‌های مرزی باید الزامی باشد؛ حکمرانی‌ای که صدور مجوز، بازرسی، گزارش‌دهی حوادث و مجازات‌های الزام‌آور برای ساخت یا انتشار غیرمجاز سامانه‌های خطرناک را دربرگیرد. معاهده‌ای که سازوکار راستی‌آزمایی نداشته باشد، چیزی جز نمایش دیپلماتیک نیست.

برخی معتقدند هشدارهای شدید درباره ابرهوش، به‌جای ایجاد احساس مسئولیت، ممکن است به درماندگی و فلج‌شدن اراده عمومی بینجامد. چگونه می‌توان درباره خطرهایی چنین عظیم سخن گفت، بی‌آنکه مردم احساس کنند آینده از پیش از دست رفته است؟

هشدارهای صریح برای فلج کردن مردم نیستند؛ هدف آن‌ها جلوگیری از حرکت ناآگاهانه و بی‌فکر به سوی فاجعه است. بشریت پیشتر نیز، هنگامی که خطرها آشکار شده‌اند، مسیر خود را تغییر داده است؛ از کنترل تسلیحات هسته‌ای و مقابله با تخریب لایه اُزن گرفته تا امنیت زیستی، ایمنی هوانوردی و سلامت عمومی. واکنش عاطفی درست، ناامیدی نیست؛ بلکه جدی گرفتن مسئله است. مردم باید بدانند که آینده نه به‌طور خودکار امن خواهد بود و نه به‌طور خودکار از دست رفته است. بدترین پیام، اطمینان‌بخشی دروغین است و دومین پیام بد، القای درماندگی و ناتوانی. پیام درست این است: ما با یک وضعیت اضطراری روبه‌رو هستیم و شرایط اضطراری، اقداماتی متناسب با بزرگی مخاطرات و پیامدهای آن می‌طلبد.

اگر قرار باشد جوانان، دانشجویان و مخاطبان غیرمتخصص تنها یک پیام از کتاب شما دریافت کنند، آن پیام چه باید باشد: ترس از هوش مصنوعی، کنترل آن، فهم آن، یا بازاندیشی در معنای انسان‌بودن پیش از آنکه چیزی باهوش‌تر از خودمان بسازیم؟

پیام کتاب صرفاً ترسیدن از هوش مصنوعی نیست؛ زیرا ترس بدون فهم، فایده‌ای ندارد. پیام اصلی این است که خلق چیزی باهوش‌تر از کل بشریت، صرفاً یک پروژه مهندسی دیگر نیست، بلکه تصمیمی درباره این است که در آینده، اختیار و زمام جهان در دست چه کسی یا چه چیزی خواهد بود. پیش از آنکه ذهن‌هایی توانمندتر از ذهن خود بسازیم، باید از خود بپرسیم آیا می‌توانیم آن‌ها را بفهمیم، با ارزش‌ها و اهداف انسانی هم‌راستا کنیم، تحت کنترل نگه داریم و بدون واگذاری عاملیت و اختیار انسان، در کنار شان زندگی کنیم؟ اگر نتوانیم به این پرسش‌ها پاسخ دهیم، اقدام مسئولانه شتاب‌بخشیدن به توسعه نیست، بلکه خویشتن‌داری و محدودکردن آن است. هوش قدرت آفرین است، اما خرد یعنی بدانیم چه زمانی نباید قدرتی را خلق کنیم که توان کنترل آن را نداریم.