ماشین می‌گوید «دزدی نمی‌کنم»؛ اما می‌تواند دزدی کند!

درگفت‌وگو با پروفسور کالین آلن به این پرسش پرداختیم که آیا ساختن ماشین‌های اخلاقی مهم است یا باید نگران استفاده‌کنندگان از این ماشین‌ها بود



هوش مصنوعی دیگر فقط مسئله‌ای فنی درباره قدرت محاسبه ماشین‌ها نیست؛ هر چه بیشتر وارد حوزه تصمیم‌گیری می‌شود، پرسش‌های اخلاقی جدی‌تری نیز پیش‌روی ما می‌گذارد: آیا ماشین می‌تواند واقعاً میان درست و غلط تمایز بگذارد؟ چه کسی باید ارزش‌های اخلاقی آن را تعیین کند؟ و اگر یک سامانه هوشمند تصمیمی نادرست یا زیان‌بار گرفت، مسئولیت با چه کسی است؟



محمدجواد استادی
Info@ostadiofficial.ir
پژوهشگر مطالعات فرهنگی

آیا در حال نزدیک‌شدن به ماشین‌هایی هستیم که واقعاً اخلاقی‌اند، یا صرفاً ماشین‌هایی می‌سازیم که در شبیه‌سازی رفتار اخلاقی روزبه‌روز ماهرتر می‌شوند؟

ما در صفحه ۲۰۲ کتاب می‌نویسیم: «نگرانی درباره این که آیا سامانه‌های مصنوعی می‌توانند عامل اخلاقی «واقعی» باشند، ما را از پرسش‌های مهم‌تر درباره چگونگی طراحی سامانه‌هایی که در موقعیت‌های اخلاقاً حساس رفتار مناسبی داشته باشند، منحرف می‌کند.» ما میان عاملیت اخلاقی عملیاتی، عاملیت اخلاقی کارکردی و عاملیت اخلاقی کامل تمایز قائل می‌شویم. عاملیت اخلاقی عملیاتی زمانی است که برنامه‌نویس ماشین، همه موقعیت‌های دارای اهمیت اخلاقی را که ماشین ممکن است در آن‌ها قرار بگیرد از پیش پیش‌بینی کرده و دقیقاً رفتاری را که در هریک از آن موقعیت‌ها مطلوب است برنامه‌ریزی کرده باشد. در این حالت، خود ماشین هیچ نوع محاسبه یا استدلالی درباره این که چه کاری باید انجام دهد، صورت نمی‌دهد. عاملیت اخلاقی کارکردی، در مقابل، مستلزم چنین محاسبه‌ای است؛ حتی اگر ماشین از همه ویژگی‌های عاملیت اخلاقی انسان برخوردار نباشد. برای مثال، ماشینی که برای انتخاب میان کنش‌های مختلف در موقعیت‌های گوناگون، از ترکیبی از اصول کلی اخلاقی و یک پایگاه داده از تصمیم‌های اخلاقی انسان‌ها استفاده کند، دارای عاملیت اخلاقی کارکردی است. در این جا برنامه‌نویس از قبل دقیقاً پیش‌بینی نکرده است که در یک موقعیت جدید چه باید کرد یا ماشین چه خواهد کرد. هیچ‌یک از این دو نوع عاملیت، همان عاملیت اخلاقی کامل در سطح انسانی نیست. گمان می‌کنم ممکن است شما «کامل» را معادل «واقعی» در نظر بگیرید، اما از نظر من این دقیقاً همان نوع مسئله‌ای است که حوزه مانیازی ندارد خود را با آن منحرف کند. تمرکز باید بر این باشد که ماشین‌ها بتوانند در موقعیت‌هایی که برنامه‌نویسان آن‌ها را به‌طور کامل پیش‌بینی نکرده‌اند، تصمیم درست بگیرند.

اضافه می‌کنم که این نوع شایستگی می‌تواند واقعی باشد، حتی اگر با همه ظرفیت‌هایی که یک انسان دارد همراه نباشد؛ همان‌طور که دانشجویان پزشکی ممکن است هنوز از نظر پزشکی کاملاً صلاحیت حرفه‌ای نداشته باشند، اما در عین حال دارای نوعی شایستگی واقعی هستند. ما استدلال کردیم که اخلاق کارکردی هدف مهندسی مناسبی است. از این منظر، روشن است که می‌توان مدل‌های زبانی بزرگ را به گونه‌ای واداشت که داوری‌های اخلاقی‌ای ارائه دهند که کمابیش با داوری‌هایی که برخی یا اکثر انسان‌ها انجام می‌دهند، مطابقت داشته باشد. اما واداشتن مدل‌های زبانی بزرگ به این که فقط رفتارهای اخلاقاً مناسب تولید کنند، مستلزم ایجاد سازوکارهای پیچیده و گسترده‌ای در اطراف هسته اصلی این مدل‌هاست. بدون این حفاظ‌ها، به همان اندازه آسان است که با شیوه‌های خاصی از آن‌ها پاسخ‌هایی آشکارا نامناسب گرفت. حتی با وجود حفاظ‌هایی که از طریق یادگیری تقویتی از بازخورد انسانی یا RLHF ایجاد شده‌اند، یا با قرار دادن نوعی پوسته یا «مهار» در اطراف سامانه، باز هم می‌توان مدل‌های زبانی را واداشت که کارهای نامناسب انجام دهند؛ چه از طریق صورت‌بندی خاص درخواست‌ها و چه از طریق سپردن وظایفی به آن‌ها که به‌سادگی و بدون انجام رفتارهای نامناسب قابل انجام نیستند. بنابراین می‌توانم بگویم هنوز فاصله زیادی داریم تا ماشین‌ها بتوانند همانند اکثر انسان‌ها خودتنظیمی داشته باشند؛ هرچند انسان‌ها نیز غالباً در این زمینه شکست می‌خورند. سطح شبیه‌سازی قطعاً بسیار بالاتر از پنج سال پیش است، اما هنوز راه درازی در پیش داریم.

عنوان فرعی کتاب شما از «آموزش درست و غلط به ربات‌ها» سخن می‌گوید. اما پس از هزاران سال فلسفه، دین، حقوق و تجربه تاریخی، انسان‌ها همچنان اختلاف‌های عمیقی درباره درست و غلط دارند. در این صورت دقیقاً چه چیزی را باید به ماشین‌ها بیاموزیم: حقیقت اخلاقی، اجماع اجتماعی، قوانین موجود یا ترجیحات کسانی که آن‌ها را طراحی و تأمین مالی می‌کنند؟

«آموزش درست و غلط به ربات‌ها» در وهله نخست عنوان خوبی است، چون در زبان انگلیسی نوعی هم‌آوایی جذاب دارد! اما در عین حال به‌شکلی ظریف به یک تنش موجود در آموزش اخلاقی اشاره می‌کند: آیا آموزش اخلاقی یعنی به انسان‌ها ـ و اکنون به ماشین‌ها ـ یک آیین‌نامه رفتاری بیاموزیم؟ یا به این معناست که به انسان‌ها ـ و شاید ماشین‌ها ـ یاد بدهیم چگونه درباره مسائل اخلاقی فکر و استدلال کنند؟

کالین آلن، فیلسوف شناخته‌شده علوم شناختی و هوش مصنوعی، از جمله متفکرانی است که سال‌ها پیش از فراگیرشدن مدل‌های زبانی بزرگ به این پرسش‌ها پرداخته است. او استاد ممتاز فلسفه و رئیس دپارتمان فلسفه دانشگاه کالیفرنیا، سانتا باربارا است و پژوهش‌هایش بر فلسفه ذهن، شناخت، هوش مصنوعی و شناخت حیوانی متمرکز است

سنت فلسفه سکولار قطعاً بر رویکرد دوم تأکید دارد، در حالی که سنت‌های دینی ــ و همچنین چیزهایی مانند آیین‌نامه‌های حرفه‌ای ــ بیشتر به رویکرد نخست گرایش دارند. با این حال، رویکرد مبتنی بر آیین‌نامه رفتاری بسیار محدود است. «قتل نکن» اصل ساده و خوبی به نظر می‌رسد، اما هیچ توضیحی نمی‌دهد که چه چیزی را نباید کشت: انسان‌های دیگر؟ انسان‌هایی که در صدد کشتن شما نیستند؟ حیوانات، از جمله حشرات؟ گیاهان؟ باکتری‌ها؟ حتی برای این که بفهمیم این قاعده دقیقاً چه معنایی دارد، به کار مفهومی بسیار زیادی نیاز است. به همین دلیل فلسفه سکولار تمرکز خود را بر آموزش شیوه استدلال می‌گذارد، نه بر این که افراد باید به چه نتیجه‌ای برسند. در این چارچوب نیز جایی برای اختلاف‌نظر معقول وجود دارد. تقریباً هر کسی که درباره این مسئله فکر کند، موافق است که کشتن یک انسان دیگر صرفاً برای تفریح کار نادرستی است. بسیاری نیز موافق‌اند که کشتن حیوان برای تفریح نادرست است، اما حتی گیاه‌خوارانی را می‌شناسم که معتقدند شکار گوزن می‌تواند راه قابل قبولی برای کنترل افزایش بیش از حد جمعیت و حفاظت از زیست‌بوم باشد. بنابراین بله، انسان‌ها همچنان درباره درست و غلط اختلاف دارند و این اختلاف‌ها زمانی خطرناک می‌شوند که یک چارچوب اخلاقی، دینی یا سیاسی واحد به شکلی جزمی بر دیگران تحمیل شود. پس درباره ماشین‌ها نیز نباید بخواهیم آن‌ها صرفاً بازتاب‌دهنده ارزش‌های سازندگان خود باشند. باید بخواهیم ماشین‌ها درون یک چارچوب مدار اگر و کثرت‌گرا به بهتر شدن استدلال ما کمک کنند؛ چارچوبی که بپذیرد درباره موارد ساده توافق گسترده‌ای وجود دارد ـ مثلاً این که انسان‌های دیگر را برای تفریح نکشیم یا از حساب‌های بانکی آنلاین پول ندزدیم ـ اما درباره موارد دشوار، امکان اختلاف‌نظر وجود دارد.

شما میان ماشینی که اخلاقی رفتار می‌کند و ماشینی که واقعاً می‌توان آن را «عامل اخلاقی» دانست، تمایز می‌گذارید. آیا ماشینی که فاقد آگاهی، قصد، احساس گناه و توانایی رنج‌کشیدن است، می‌تواند واقعاً اخلاقی باشد؟ یا چنین ماشینی صرفاً ابزاری بسیار پیچیده برای اجرای تصمیم‌های اخلاقی انسان باقی خواهد ماند؟

بله. آگاهی، عواطف و مواردی از این دست در مرکز دیدگاه شخصی من این است که باید سلاح‌های خودمختار را ممنوع کنیم. اما فکر می‌کنم برای تحقق چنین ممنوعیتی احتمالاً باید نوعی فاجعه رخ دهد، زیرا توسعه این نوع سلاح‌ها از هم‌اکنون به‌طور جدی در جریان است.

کتاب شما مسئله «حقوق ماشین» را نیز مطرح می‌کند. اگر یک سامانه مصنوعی در آینده نوعی آگاهی، تجربه یا عاملیت اخلاقی پیدا کند، آیا خاموش‌کردن آن، پاک‌کردن حافظه‌اش یا مجبورکردن آن به کار بی‌پایان می‌تواند شکلی از آسیب یا استثمار محسوب شود؟ چگونه می‌توان تشخیص داد که مطالبه حقوق برای ماشین‌ها واقعاً ناشی از یک ضرورت اخلاقی است، نه حاصل تمایل ما به فریب‌خوردن در برابر رفتارهای انسان‌وار ماشین؟

فکر می‌کنم حالت دوم درست است و همین مسئله پشت حرف قبلی من قرار داشت که گفتم پرسش درباره عاملیت اخلاقی «واقعی»، ما را از مشکلات فوری‌تر منحرف می‌کند؛ مشکلات مربوط به این که چگونه با فناوری‌ای که واقعاً قادر به ساخت آن هستیم، بهترین کاری را که می‌توانیم انجام دهیم.

شاید بنیادی‌ترین نقد این باشد که چالش اصلی ما ساختن «ماشین‌های اخلاقی» نیست، بلکه ایجاد نهادها و جوامعی است که در آن‌ها انسان‌ها از ماشین‌ها مسئولانه استفاده کنند. آیا تلاش برای دادن «وجدان» به ربات‌ها این خطر را ندارد که توجه ما را از قدرت شرکت‌ها، دولت‌ها و سازمان‌های نظامی‌ای که این ماشین‌ها را می‌سازند و به کار می‌گیرند منحرف کند؟ اگر امروز می‌خواستید عنوان فرعی تازه‌ای برای کتاب انتخاب کنید، همچنان می‌نوشتید «آموزش درست و غلط به ربات‌ها»، یا در عوض می‌پرسیدید: چگونه می‌توانیم به انسان‌ها بیاموزیم که داوری اخلاقی خود را به ماشین‌ها واگذار نکنند؟

اگر امروز می‌خواستم این کتاب را بنویسم، قطعاً مطالب بیشتری درباره زمینه‌های اجتماعی ـ فنی‌ای می‌نوشتم که عوامل اخلاقی مصنوعی ممکن است درون آن‌ها قرار بگیرند. اما فکر نمی‌کنم از نوشتن کتابی که نوشتم پشیمان باشم. تمرکز آن کتاب بر کاوش در این مسئله بود که خود ماشین ممکن است از چه ظرفیت‌هایی برخوردار شود و من همچنان فکر می‌کنم تمرکز بر این پرسش، پرسشی کاملاً مشروع است.

کالین آلن، فیلسوف شناخته‌شده علوم شناختی و هوش مصنوعی، از جمله متفکرانی است که سال‌ها پیش از فراگیرشدن مدل‌های زبانی بزرگ به این پرسش‌ها پرداخته است. او استاد ممتاز فلسفه و رئیس دپارتمان فلسفه دانشگاه کالیفرنیا، سانتا باربارا است و پژوهش‌هایش بر فلسفه ذهن، شناخت، هوش مصنوعی و شناخت حیوانی متمرکز است

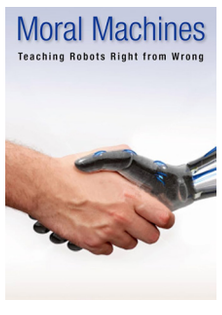
آن چیزی قرار دارند که ما در فصل دهم کتاب «قوای فراعقلانی» می‌نامیم و فرض ما این است که این‌ها به نحوی برای برخورداری از عاملیت اخلاقی کامل ضروری‌اند. از نظر من، ماشین‌های کنونی دارای تجربه آگاهانه نیستند و با توجه به این که از نظر علمی درباره منشأ آگاهی خودمان نیز اطلاعات بسیار اندکی داریم، نمی‌دانیم چگونه می‌توانیم ـ یا حتی آیا اصلاً می‌توانیم ـ ماشین‌هایی بسازیم که چنین تجربه‌هایی داشته باشند. همکار من معتقد است احتمالاً قادر به انجام چنین کاری نیستیم، اما من نسبت به این امکان دیدگاه بازتری دارم. با این حال، فکر می‌کنم رویکردهای کنونی، از جمله مدل‌های زبانی بزرگ، برای این کار کافی نیستند. همچنین نیازمند نظریه‌پردازی بیشتری درباره این مسئله هستیم که چگونه عواطف و دیگر حالات آگاهانه، صرفاً عوارض جانبی زیست‌شناسی ما نیستند، بلکه از نظر کارکردی در شکل‌دادن به شناخت انسانی ـ از جمله شناخت اخلاقی ما ـ نقش مهمی دارند. در حال حاضر یک مدل زبانی بزرگ می‌تواند عصبانی بودن یا خوشحال بودن را شبیه‌سازی کند؛ اما حالتی که در آن لحظه در آن قرار دارد، آن‌طیف گسترده از تأثیراتی را که خشم یا شادی بر رفتار و پردازش اطلاعات در انسان می‌گذارند، بر آن مدل ندارد. تا زمانی که این وجه از عاملیت اخلاقی در ماشین‌ها توسعه نیافته باشد، باید همواره آگاه باشیم که با یک «ابزار بسیار پیچیده» مواجهیم ـ همان تعبیری که شما به کار بردید ـ و نباید دچار این اشتباه شویم که گمان کنیم با موجودی کاملاً معادل خودمان روبه‌روهستیم.

شما نشان می‌دهید که نظریه‌های کلاسیک اخلاقی ـ از فایده‌گرایی تا وظیفه‌گرایی ـ به‌سادگی قابل تبدیل‌شدن به الگوریتم نیستند. آیا این دشواری عمدتاً ناشی از محدودیت‌های محاسباتی ماشین‌هاست، یا نشان می‌دهد نظریه‌های اخلاقی خود ما نیز برای هدایت زندگی پیچیده انسانی ناکافی‌اند؟ آیا اکنون اخلاق فضیلت، اخلاق مراقبت یا نوعی کثرت‌گرایی اخلاقی را برای طراحی ماشین‌های اخلاقی مناسب‌تر می‌دانید؟

در جریان کار روی کتاب، متوجه شدم که همدلی بیشتری با خاص‌گرایی اخلاقی پیدا کرده‌ام؛ دیدگاهی که معتقد است هر موقعیت دارای ویژگی‌های منحصربه‌فردی است که یک قاعده کلی نمی‌تواند به اندازه کافی آن‌ها را پوشش دهد. البته من تا آن جا که برخی خاص‌گرایان پیش می‌روند، نمی‌روم. برای مثال، راسل دنسی معتقد است نظریه‌های اخلاقی بی‌فایده‌اند. من بیشتر تمایل دارم نظریه‌ها را نوعی راهنمای اکتشافی برای استدلال خودبدانم؛ چیزی شبیه قیودنرم که وارد یک نظام موازنه می‌شوند؛ نظامی که در آن نمی‌توان همه مطلوب‌های رقیب را به‌طور همزمان برآورده کرد. اما جزئیات موقعیت‌های خاص به یک فرد با تجربه ـ یا خردمند ـ اجازه می‌دهد نوعی راه‌حل خلاقانه پیدا کند که افراد بتوانند بر سر این مسئله توافق کنند که احتمالاً بهترین کاری است که در آن موقعیت می‌توان انجام داد.

یکی از نقدهای مهم به کتاب این است که اخلاق انسانی را بیش از اندازه آسان به‌عنوان الگوی اخلاق ماشین در نظر می‌گیرد. برخی منتقدان معتقدند ماشین‌های تخصصی شاید به شکلی محدود، متناسب با وظیفه و اساساً غیرانسانی از اخلاق نیاز داشته باشند. برای مثال، آیا یک خودروی خودران واقعاً به چیزی شبیه وجدان انسانی نیاز دارد؟ یا قواعد دقیق ایمنی، سوابق داده‌ای شفاف و خطوط مسئولیت روشن کفایت می‌کنند؟ این نقد تا چه اندازه موضع شما را تغییر داده است؟

این نقد تغییر زیادی در موضع من ایجاد نکرده است. همان‌طور که در کتاب نیز بحث کردیم، سامانه‌هایی که برای کاربردهای خاص ساخته می‌شوند، فقط به همان میزانی از اخلاق کارکردی نیاز دارند که برای آن موقعیت مناسب است. مشکل قواعد دقیق ایمنی ـ که شکلی از اخلاق عملیاتی هستند ـ این است که مستلزم آن هستند که برنامه‌نویس همه موقعیت‌هایی را که ماشین ممکن است در آن‌ها قرار بگیرد، از پیش پیش‌بینی کند.



می‌توان تلاش کرد برای خودروهای خودران قواعد دقیق ایمنی نوشت؛ اما فرض کنید این قاعده می‌گوید خودرو همیشه باید در فاصله معینی از عابر پیاده‌ای که در مسیر قرار گرفته متوقف شود. حال اگر آن فرد حاضر نباشد از جایش حرکت کند چه باید کرد؟ هیچ کاری نکنیم، حتی اگر باعث ایجاد ترافیک شود؟ برای او بوق بزنیم؟ مگر این که در محدوده‌ای باشیم که بوق‌زدن ممنوع است! کمی به او نزدیک شویم تا وادار شود کنار برود؟ یا خودرو را به شکلی تهاجمی به سمت او حرکت دهیم تا از جاده بترسد و کنار برود؟ همه این گزینه‌ها مستلزم ارزیابی انعطاف‌پذیر موقعیت‌هایی هستند که شاید برنامه‌نویسان آن‌ها را به‌طور کامل پیش‌بینی نکرده باشند و نمی‌توان با قواعد ایمنی دقیقی مانند «همیشه فاصله مشخصی را حفظ کن» آن‌ها را حل کرد. حتی واگذار کردن کنترل به یک راننده انسانی در چنین شرایطی نیز به‌طور کامل از پیامدهای بدی مانند تصادف یا ایجاد ترافیک جلوگیری نمی‌کند.

خود ماشین هیچ نوع محاسبه یا استدلالی درباره این که چه کاری باید انجام دهد، صورت نمی‌دهد. عاملیت اخلاقی کارکردی، در مقابل، مستلزم چنین محاسبه‌ای است؛ حتی اگر ماشین از همه ویژگی‌های عاملیت اخلاقی انسان برخوردار نباشد

بخش بزرگی از اخلاق انسانی با عاطفه، همدلی، آسیب‌پذیری، بدن‌مندی و تجربه زیسته پیوند دارد. آیا هوش مصنوعی‌ای که بدن ندارد و قادر به رنج‌کشیدن نیست، می‌تواند واقعاً تحقیر، ترس، فقر یا مرگ را درک کند؟ آیا شبیه‌سازی همدلی برای رفتار اخلاقی کفایت می‌کند، یا همدلی بدون احساس صرفاً اجرایی متقاعدکننده است؟

قبلاً در ارتباط با آنچه «قوای فراعقلانی» نامیدیم، به این مسئله اشاره کردم. همچنین در برابر پرسش‌هایی درباره نسخه «واقعی» این ظرفیت‌ها مقاومت کرده‌ام، زیرا آن‌ها را عاملی برای منحرف‌شدن از چالش‌های مهندسی‌ای می‌دانم که امکان پرداختن عملی‌تری به آن‌ها وجود دارد و معتقدم فیلسوفان باید در مواجهه با همین چالش‌ها به مهندسان کمک کنند. همدلی شبیه‌سازی‌شده شاید برای همه رفتارهای اخلاقی کافی نباشد، اما ممکن است برای برخی از آن‌ها کافی باشد. صرفاً ارائه یک اجرای متقاعدکننده در بعضی موقعیت‌ها ممکن است کار درستی باشد و در موقعیت‌هایی دیگر کارنادرستی. فیلسوفان می‌توانند به مهندسان کمک کنند تشخیص دهند چه چیزی میان این دو وضعیت تفاوت ایجاد می‌کند.

ایده اخلاق ماشین بلافاصله یک پرسش سیاسی را پیش می‌کشد: اخلاقِ چه کسی؟ قدرتمندترین سامانه‌های هوشِ مصنوعی جهان توسط تعداد اندکی شرکت ساخته می‌شوند که عمدتاً در ایالات متحده و چند کشور دیگر مستقرند. چگونه می‌توان مانع شد که ارزش‌های فرهنگی و سیاسی یک اقلیت فناورانه، تحت عنوان «اخلاق جهانی» بر میلیاردها انسان تحمیل شود؟

این نوع پرسش سیاسی شاید تا حدی خارج از حوزه تخصص من باشد. من واقعاً نمی‌دانم چگونه می‌توان قدرت شرکت‌های بزرگ را مهار کرد؛ چه شرکت‌های هوش مصنوعی باشند، چه شرکت‌های نفتی، بانک‌ها یا دیگران. همه آن‌ها ارزش‌های خود را بر میلیاردها انسان تحمیل می‌کنند و مقیاس فعالیت‌هایشان از حوزه نفوذ هر کشور منفردی ــ حتی ثروتمندترین و قدرتمندترین کشورها ـ فراتر می‌رود. شاید رونق کنونی هوش مصنوعی از این نمونه‌های دیگر نگران‌کننده‌تر باشد؛ زیرا می‌توانیم اخلاق فناوری را به شیوه‌هایی مورد پرسش قرار دهیم که درباره محصولات یا خدمات حوزه‌های دیگر چندان امکان‌پذیر نیست. اما شاید همین موضوع دلیلی برای خوش‌بینی باشد. کار بر ان نهایی بسیار آسان‌تر می‌توانند مشکلات و نقص‌های مدل‌های زبانی بزرگ را کشف و آشکار کنند و این موضوع شرکت‌ها را مجبور می‌کند این مشکلات را به رسمیت بشناسند و شاید برای حل آن‌ها کاری انجام دهند. به همین دلیل است که اخیراً اخبار و توجه زیادی درباره استخدام متخصصان اخلاق توسط شرکت‌هایی مانند Anthropic و OpenAI مطرح شده است. این که این اقدامات تا چه اندازه مؤثر خواهند بود، هنوز باید در عمل مشخص شود.

ماشین‌هایی که اخلاق را از داده‌های تولیدشده توسط انسان یاد می‌گیرند، ناگزیر با تاریخی مملو از تبعیض، استعمار، نژادپرستی، نابرابری طبقاتی و خشونت مواجه می‌شوند. یک ماشین چگونه می‌تواند میان آنچه انسان‌ها معمولاً انجام داده‌اند و آنچه اخلاقاً باید انجام دهند تمایز بگذارد؟ آیا آموزش اخلاق از داده‌های تاریخی این خطر را ندارد که پیش‌داوری‌های اکثریت را به «وجدان الگوریتمی» تبدیل کند؟

بله؛ این دقیقاً مشکل رویکردهای صرفاً پایین به بالاست. اما همان‌طور که توضیح دادم، یک سامانه ترکیبی که به‌خوبی طراحی شده باشد، ابزارهایی در اختیار دارد که هم رویه‌های عملی و هم نظریه‌ها را نقد کند. مدل‌های زبانی بزرگ کنونی از چنین قابلیتی برخوردار نیستند، مگر در حدی که بر اساس بحث‌های پیشین متخصصان اخلاق آموزش دیده باشند و همان‌گونه که قبلاً نیز اشاره کردم، گرچه مدل‌های زبانی بزرگ می‌توانند نظریه‌ها را گزارش و بازتولید کنند، همچنان این مشکل وجود دارد که اصول نظری همیشه به‌طور مؤثر رفتار آن‌ها را کنترل نمی‌کنند. بنابراین در این فاصله باید نه‌فقط بر ارتقای توانایی ماشین‌ها، بلکه بر بهبود سامانه‌های اجتماعی ـ فنی که این ماشین‌ها درون آن‌ها قرار می‌گیرند تمرکز کنیم؛ از جمله ایجاد سازوکارهایی برای نظارت انسانی تا در برابر بازتولید بی‌فکرانه سوگیری‌ها از طریق الگوریتم‌ها حفاظ ایجاد شود.

برخی پژوهشگران هشدار می‌دهند که ماشین‌های اخلاقی ممکن است ارزش‌های عصر خود ما را منجمد کنند و مانعی در برابر پیشرفت اخلاقی آینده شوند. از سوی دیگر، ماشینی که دائم ارزش‌های خود را تغییر می‌دهد، ممکن است غیرقابل پیش‌بینی و خطرناک شود. چگونه می‌توان سامانه‌ای طراحی کرد که از نظر اخلاقی اصلاح‌پذیر باقی بماند، بی‌آنکه هر بار افکار عمومی، سیاست دولت یا منافع شرکت‌ها تغییر کرد، به موجودی اخلاقاً فرصت‌طلب تبدیل شود؟

باز هم پاسخ این است که باید بر سامانه اجتماعی ـ فنی تمرکز کنیم، نه صرفاً بر عامل اخلاقی مصنوعی. فکر می‌کنم یکی از نقدهای موجهی که به کتاب Moral Machines وارد شد این بود که کتاب بیش از اندازه بر خود عامل اخلاقی مصنوعی تمرکز کرده بود. کارهای جدیدتر من در این حوزه ـ مانند پروژه «خردماشینی» که همراه با برت کارلن انجام می‌دهم ــ در تلاش است پرسش‌هایی مربوط به سامانه اجتماعی ـ فنی را بیشتر مورد توجه قرار دهد.

آیا هرگز باید به یک سامانه غیرشفاف اجازه داد درباره زندگی انسان‌ها و حقوق بنیادین آن‌ها تصمیم بگیرد، حتی اگر از نظر آماری عملکردش از تصمیم‌گیرندگان انسانی بهتر باشد؟

فکر می‌کنم واژه کلیدی در این پرسش «غیرشفاف» است و همین واژه پاسخ منفی را تعیین می‌کند. اگر ندانیم چرا ماشین تصمیم خاصی می‌گیرد، نباید از آن برای جایگزین‌کردن یک تصمیم‌گیرنده انسانی استفاده کنیم. ممکن است افراد بخواهند با آن مشورت کنند، اما حتی در این صورت نیز باید با شک و احتیاط با آن برخورد کنند. بخشی از استدلالی که من و کارلن مطرح می‌کنیم این است که برخورد خردمندانه با هوش مصنوعی مستلزم این است که نه‌فقط بدانیم چه کارهایی می‌تواند انجام دهد، بلکه بفهمیم حالت‌ها و الگوهای شکست آن چیست. تنها زمانی که درک خوبی از این داشته باشیم که سامانه چه‌وقت و چرا شکست می‌خورد، می‌توانیم مسئولانه از آن در موقعیت‌هایی استفاده کنیم که ممکن است شرایط پیش‌بینی‌نشده‌ای در آن‌ها رخ دهد.

استفاده از هوش مصنوعی در جنگ، پرسش مرکزی کتاب شما را به شکلی بسیار هولناک‌تر مطرح می‌کند: آیا هرگز می‌توان به یک ماشین از نظر اخلاقی اجازه داد به‌طور مستقل تصمیم بگیرد انسانی کشته شود؟ آیا تلاش برای توسعه «سلاح‌های اخلاقی» خطر مشروعیت‌بخشیدن به سامانه‌های تسلیحاتی خودمختار را در پی ندارد؟ از نظر شما چه خط قرمز اخلاقی‌ای هرگز نباید شکسته شود، حتی اگر ادعا شود ماشین‌ها از سربازان انسانی دقیق‌ترند؟

دیدگاه شخصی من این است که باید سلاح‌های خودمختار را ممنوع کنیم. اما فکر می‌کنم برای تحقق چنین ممنوعیتی احتمالاً باید نوعی فاجعه رخ دهد، زیرا توسعه این نوع سلاح‌ها از هم‌اکنون به‌طور جدی در جریان است.

کتاب شما مسئله «حقوق ماشین» را نیز مطرح می‌کند. اگر یک سامانه مصنوعی در آینده نوعی آگاهی، تجربه یا عاملیت اخلاقی پیدا کند، آیا خاموش‌کردن آن، پاک‌کردن حافظه‌اش یا مجبورکردن آن به کار بی‌پایان می‌تواند شکلی از آسیب یا استثمار محسوب شود؟ چگونه می‌توان تشخیص داد که مطالبه حقوق برای ماشین‌ها واقعاً ناشی از یک ضرورت اخلاقی است، نه حاصل تمایل ما به فریب‌خوردن در برابر رفتارهای انسان‌وار ماشین؟

فکر می‌کنم حالت دوم درست است و همین مسئله پشت حرف قبلی من قرار داشت که گفتم پرسش درباره عاملیت اخلاقی «واقعی»، ما را از مشکلات فوری‌تر منحرف می‌کند؛ مشکلات مربوط به این که چگونه با فناوری‌ای که واقعاً قادر به ساخت آن هستیم، بهترین کاری را که می‌توانیم انجام دهیم.

شاید بنیادی‌ترین نقد این باشد که چالش اصلی ما ساختن «ماشین‌های اخلاقی» نیست، بلکه ایجاد نهادها و جوامعی است که در آن‌ها انسان‌ها از ماشین‌ها مسئولانه استفاده کنند. آیا تلاش برای دادن «وجدان» به ربات‌ها این خطر را ندارد که توجه ما را از قدرت شرکت‌ها، دولت‌ها و سازمان‌های نظامی‌ای که این ماشین‌ها را می‌سازند و به کار می‌گیرند منحرف کند؟ اگر امروز می‌خواستید عنوان فرعی تازه‌ای برای کتاب انتخاب کنید، همچنان می‌نوشتید «آموزش درست و غلط به ربات‌ها»، یا در عوض می‌پرسیدید: چگونه می‌توانیم به انسان‌ها بیاموزیم که داوری اخلاقی خود را به ماشین‌ها واگذار نکنند؟

اگر امروز می‌خواستم این کتاب را بنویسم، قطعاً مطالب بیشتری درباره زمینه‌های اجتماعی ـ فنی‌ای می‌نوشتم که عوامل اخلاقی مصنوعی ممکن است درون آن‌ها قرار بگیرند. اما فکر نمی‌کنم از نوشتن کتابی که نوشتم پشیمان باشم. تمرکز آن کتاب بر کاوش در این مسئله بود که خود ماشین ممکن است از چه ظرفیت‌هایی برخوردار شود و من همچنان فکر می‌کنم تمرکز بر این پرسش، پرسشی کاملاً مشروع است.