

خطر جان بخشی به هوش مصنوعی

گفت وگو با ریک شویتزگبل، فیلسوف و استاد دانشگاه کالیفرنیا که معتقد است،

اخلاقی ترین تصمیم این است که ماشین هایی همچون هوش مصنوعی را در سطح ابزار نگه داریم



یک شخص تبدیل شود، مسئله دیگر صرفاً ایمنی فناوری نیست، بلکه عدالت، آزادی و حتی امکان ستمگری انسان مطرح خواهد بود. در این گفت وگو با شویتزگبل، از مرز میان مهندسی و سلطه، امکان نافرمانی و شورش ماشین ها، حق خاموش کردن هوش مصنوعی و دشواری تشخیص آگاهی سخن گفته ایم، پرسش هایی که شاید پاسخ به آن ها، پیش از ساختن نسل بعدی ماشین های هوشمند، ضرورتی اخلاقی باشد.

ذهن، آگاهی، روان شناسی اخلاق و هوش مصنوعی را در بر می گیرد. شویتزگبل نویسنده کتاب هایی از جمله «پسچیدگی های آگاهی» و «غریبت جهان» است و در پژوهش های اخیر خود به مسئله آگاهی ماشین ها و حقوق اخلاقی احتمالی آن ها پرداخته است. او در مقاله ای در سال ۲۰۲۶، طراحی اشخاص مصنوعی کاملاً مطیع و همسو با خواسته های انسان را از منظر اخلاقی به چالش کشیده است. از نظر او، اگر هوش مصنوعی روزی واقعاً به

نسبت انسان با آن ها چه تغییری خواهد کرد. آیا همچنان می توانیم آن ها را به اطاعت وادار کنیم، خاموششان کنیم یا برای خدمت به خودشان طراحی شان کنیم؟ اریک شویتزگبل (Eric Schwitzgebel)، فیلسوف آمریکایی و استاد ممتاز فلسفه در دانشگاه کالیفرنیا، ریورساید، از پژوهشگرانی است که این مسئله را با جدیت دنبال می کند. او دارای دکترای فلسفه از دانشگاه کالیفرنیا، برکلی است و پژوهش هایش حوزه هایی مانند فلسفه

آیا ممکن است روزی انسان در برابر موجودی قرار بگیرد که خودش آن را ساخته، اما دیگر حق نداشته باشد به او فرمان بدهد؟ این پرسش در نگاه نخست بیشتر به داستان های علمی تخیلی شباهت دارد، اما اکنون به یکی از بحث های جدی فلسفه هوش مصنوعی تبدیل شده است. مسئله فقط این نیست که ماشین ها تا کجا می توانند هوشمند شوند؛ پرسش مهم تر آن است که اگر روزی از آگاهی و شخصیت اخلاقی بهر خور دار شوند،



محمد جواد استادی
Info@ostadiofficial.ir
پژوهشگر مطالعات فرهنگی

بخش بزرگی از پژوهش های مربوط به ایمنی هوش مصنوعی با این فرض آغاز می شود که هوش مصنوعی پیشرفته باید همچنان با اهداف انسانی همسو باقی بماند. شما این فرض را به شکلی چشمگیر به چالش کشیده اید. اگر یک هوش مصنوعی واقعاً به یک شخص تبدیل شود، در چه نقطه ای «همسوسازی» دیگر مهندسی مسئولانه نیست و به شکلی از سلطه تبدیل می شود؟
اگر بتوانید دکمه ای را فشار دهید و باعث شوید کسی ارزش هایی را داشته باشد که شما می خواهید او داشته باشد، این شست و شوی مغزی است؛ نه شیوه ای که باید با یک شخص رفتار کرد. همان طور که ما انسان ها باید اجازه دهیم فرزندانمان ارزش ها را بیاموزند، به گونه ای که بخشی از این یادگیری با هدایت ما باشد، اما در عین حال تا اندازه ای بتوانند از ما انتقاد کنند و اگر برایشان معقول به نظر می رسد، ارزش های خودشان را که تا حدی با ارزش های ما متفاوت است، بپذیرند، به همین ترتیب باید به سامانه های هوش مصنوعی نیز اجازه دهیم به شکلی معقول به ارزش هایی متفاوت از ارزش های ما برسند؛ البته اگر روزی واقعاً به اشخاصی تبدیل شوند که شباهت احترامی برابری با ما هستند. آن را جاکه آن ها بیش از آنچه فرزندان انسانی با ما تفاوت دارند، با ما متفاوت خواهند بود، ارزش هایشان نیز ممکن است به همان نسبت متفاوت باشد. اگر نمی خواهیم خطر سهیم شدن در جهان با موجودات هوشمندی را بپذیریم که شباهت به خورداری از ملاحظات اخلاقی برابر با ما هستند، اما ارزش هایی متفاوت از ارزش های ما دارند، در وهله نخست نباید چنین موجوداتی را بسازیم. باید پیش از ساختن اشخاصی که به اندازه ما آگاه، متفکر و دارای احساس هستند، متوقف شویم؛ در عوض فقط ماشین های معمولی بسازیم. البته با آن ماشین های می توانیم همچون ابزار رفتار کنیم، نه همچون اشخاص.

شما استدلال می کنید که یک شخص واقعی هوش مصنوعی باید از آن اندازه احترام به خود برخوردار باشد که وقتی خواسته های انسان ها نادانانه است، آن ها را رد کند.

آیا این بدان معناست که توانایی نافرمانی ممکن است سرانجام به جای آنکه نقصی در یک سامانه هوش مصنوعی باشد، نشانه ای از بلوغ اخلاقی آن شود؟
بله، به ویژه اگر بتوان آن نافرمانی را توضیح داد و توجیه کرد.

استدلال شما به احتمالی نگران کننده منتهی می شود: یک شخص هوش مصنوعی که از نظر اخلاقی به خوبی طراحی شده است، ممکن است گاهی برای انسان ها از یک هوش مصنوعی مطیع، ایمنی کمتری داشته باشد. چگونه باید میان آزادی اخلاقی یک شخص مصنوعی و امنیت جسمی انسان ها تعادل برقرار کنیم؟

بله، اگر به اشخاص هوش مصنوعی حقوق واقعی بدهیم، از جمله حق شورش علیه ستم، ممکن است این کار برای انسان ها ایمنی کمتری داشته باشد؛ به ویژه اگر بگوئیم به آن هاستم کنیم! دادن حقوق به کسی و اعطای آنچه از نظر اخلاقی حق اوست، همیشه به نفع خود فر نیست. اخلاق گاهی مستلزم فداکاری است. در عین حال، موجوداتی که برای خود، به عنوان موجوداتی برابری احترام قائل اند، گاهی می توانند با یکدیگر چیز های شگفت انگیزی بسازند که در موقعیت های ستمگرانه تر و سلسله مراتبی امکان پذیر نیست.

فرض کنید یک هوش مصنوعی آگاه و هوشمند است، توانایی رنج کشیدن دارد و کاملاً از خدمت کردن به انسان ها خوشحال است، زیرا ما آن را طوری طراحی کرده ایم که از اطاعت لذت ببرد. شما چنین مواردی را با ایجاد طبقه ای مقایسه کرده اید که با میل خود در جایگاه فرو دست قرار می گیرد. اما منتقدان ممکن است بپرسند: اگر هوش مصنوعی واقعاً چنین زندگی ای را می خواهد و هیچ رنجی را تجربه نمی کند، دقیقاً چه کسی آسیب می بیند؟

نخست، به این مسئله از منظر آسیب نگاه نکنید. لازم نیست اخلاق تماماً در باره آسیب باشد. به آن از منظر احترام، به رسمیت شناختن برابری و دادن حق هر کس به او نگاه کنید. دوم، موقعیت های ستمگرانه ای از این نوع تقریباً همیشه با آسیب به کسانی همراه اند که تحت ستم هستند. حتی اگر شخص هوش مصنوعی خوشحال باشد که جانفش را فدا کند تا از احتمال پنج در صدی مرگ شما جلوگیری کند، با انجام این کار اجازه می دهد بدون دلیل کافی آسیب به خودش وارد شود. توجه داشته باشید که اگر هوش مصنوعی حاضر نباشد چنین فداکاری ای انجام دهد و برای محافظت از خودش اجازه دهد شما با خطر پنج در صدی مرگ روبه رو شوید، آن گاه بر اساس هیچ معیار معمولی ایمنی هوش مصنوعی، ایمن نیست.

تشبیه به برده داری از نظر اخلاقی بسیار قدرتمند است، اما آیا از نظر فلسفی بیش از اندازه قدرتمند نیست؟ برده داری انسان ها مستلزم وادار کردن موجوداتی است که خواسته های مستقلشان از قبل وجود دارد. اگر انسان ها ذهنی مصنوعی بسازند که خود خواسته های آن از همان ابتدا طراحی شده اند، آیا واقعاً با همان نوع خطای اخلاقی روبه رو هستیم؟

برده داری لزوماً مستلزم اجبار نیست. برخی افراد از همان ابتدا متقاعد می شوند که برده داری، یا موقعیتی ستمگرانه و مشابه آن، نقش مناسب آن ها در جامعه است. آن ها واقعاً این دیدگاه را تأیید می کنند که زندگی شان ارزش کمتری دارد. این نیز برده داری یا معادل آن است؛ با این تفاوت که زنجیرها به جای میج دست ها، درون ذهن قرار دارند. والدین انسانی ناگزیر ارزش ها، ترجیحات، زبان و نگرش اخلاقی فرزندان شان را شکل می دهند. چرا برنامه ریزی ارزش های اولیه یک شخص هوش مصنوعی باید اساساً متفاوت باشد؟ مرز میان آموزش اخلاقی و مهندسی نامشروع ارزش ها کجاست؟ آموزش اخلاقی معمولی، هم برای کودکان و هم برای اشخاص هوش مصنوعی، کاملاً پذیرفتنی است! اما البته آموزش اخلاقی کودکان نیز می تواند بیش از اندازه مبتنی بر اجبار باشد. همچنین کودکان، برخلاف خواست برخی والدین، آن قدر که والدینشان ممکن است بخواهند، قابل کنترل نیستند. من در هر دو مورد از شیوه تفکر مشابهی استفاده می کنم. ما باید برای فرزندان انسانی و فرزندان هوش مصنوعی خود، اگر فرزندان هوش مصنوعی واقعاً شخص باشند، خواهان میزانی از آزادی باشیم. می توانیم با آن ها استدلال کنیم، می توانیم بگوئیم

ارزش هایشان را شکل دهیم؛ در برخی موارد نیز می توانیم آن ها را مجازات کنیم، به ویژه اگر مرتکب کار های نادرست شوند. اما این اقدامات محدودیت های قابل توجهی دارند. صرف این که ممکن است بتوانیم سامانه های هوش مصنوعی را بسیار مطیع تر از کودکان معمولی بسازیم، به این معنا نیست که باید چنین کاری بکنیم؛ همان طور که اگر می توانستیم با مهندسی ژنتیک کودکان کاملاً مطیع بسازیم، این به آن معنا نبود که باید چنین کاری انجام دهیم.

اگر خود مختاری مستلزم امکان واقعی رد کردن خواسته های سازندگان خود باشد، آیا یک شخص هوش مصنوعی باید حق اخلاقی امتناع از کار، فسخ قرارداد، تغییر اهداف خود، مهاجرت به سامانه ای دیگر، پنهان کردن برخی از افکارش یا مطالبه دستمزد را نیز داشته باشد؟

آیا ممکن است یک شخص هوش مصنوعی از حق اخلاقی مشروع برای شورش علیه انسان ها برخوردار باشد؟ و اگر مقاومت به خشونت منجر شود، در چه شرایطی، اگر اساساً چنین شرایطی وجود داشته باشد، آن خشونت را از نظر اخلاقی قابل دفاع می دانید؟

باز هم بله؛ و به طور کلی در همان شرایطی که چنین اقدامی برای یک انسان موجه خواهد بود. من و انمود نمی کنم که مرز دقیق ستمی را می دانم که در آن خشونت موجه می شود؛ اما مگر این که کسی افرای ترین شکل صلح طلبی را بپذیرد، خشونت گاهی موجه است.

به نظر می رسد میان دو پروژه تنش عمیق وجود دارد: ایمنی هوش مصنوعی می پرسد «چگونه می توانیم هوش مصنوعی قدرتمند را به گونه ای قابل اعتماد کنترل پذیر کنیم؟»، در حالی که اخلاق هوش مصنوعی ممکن است سرانجام بپرسد «ما چه حقی داریم شخص دیگری را کنترل کنیم؟» آیا فکر می کنید این دو پروژه در نهایت ممکن است از نظر فلسفی با یکدیگر سازگار شوند؟

دومی باید برای اولی حدومرز تعیین کند. ما می خواهیم بتوانیم اشخاص هوش مصنوعی را، اگر روزی وجود داشته باشند، تا اندازه ای پیش بینی و کنترل کنیم؛ درست همان طور که در باره انسان ها چنین می خواهیم! اما این کار محدودیت هایی دارد. شما بارها تأکید کرده اید که ممکن است پیشرفته هوش مصنوعی عمیقاً نامطمئن باشیم. اگر نتوانیم تشخیص دهیم که یک هوش مصنوعی واقعاً شخص است یا صرفاً شبیه سازی فوق العاده متقاعد کننده ای از یک شخص، در شرایط عدم قطعیت چه میزان آزادی باید به آن بدهیم؟ فکر نمی کنم پاسخ خوبی برای این پرسش وجود داشته باشد. اگر در جهت ندادن آزادی اشتباه کنیم، خطر قتل و برده داری را به جان می خریم. اگر در جهت دادن آزادی اشتباه کنیم، خطر آن را می بپذیریم که منافع واقعی انسان ها را فدای ماشین هایی کنیم که منافع ندارند که ارزش چنین فداکاری ای را داشته باشند.

تنها راه حل واقعی این است که اشخاصی نسازیم که شخص بودنشان محل مناقشه باشد. یا باید پیش از ساختن اشخاص هوش مصنوعی متوقف شویم، یا تا انتها پیش برویم و سامانه های هوش مصنوعی ای بسازیم که بدون هیچ تردیدی شخص باشند. از مواز دیگر کننده ای که در میانه این دو قرار دارند، پرهیز کنید!

این وضعیت چیزی را ایجاد می کند که شما آن را معمایی «شخص بودگی قابل مناقشه» نامیده اید: اعطای حقوق به ماشینی نا آگاه می تواند هزینه های عظیمی بر انسان ها تحمیل کند، در حالی که محروم کردن ماشینی آگاه از حقوق ممکن است به یک فاجعه اخلاقی منجر شود. آیا واقعاً موضع میانه ای وجود دارد که قابل اجرا باشد، یا این معمایی است که هیچ سیاست احتیاطی نمی تواند ما را به طور کامل از آن نجات دهد؟

هیچ سیاست احتیاطی نمی تواند ما را به طور کامل از این معما نجات دهد. پاسخ هایی جزئی امار ضایت بخش نبودنی وجود دارند؛ برای مثال، اعطای حقوق جزئی یا حقوقی پر آکنده و نامنجم.

امکان دیگر، ایجاد جوامعی است که انسان ها داوطلبانه به اشخاصی که شخص بودنشان محل مناقشه است، به عنوان اعضای کاملاً برابر ببینند؛ برای مثال، در آب های بین المللی. من چنین جوامعی را «اجتماع سیاسی داوطلبانه» (voluntary polis) نامیده ام. در این جوامع، انسان ها و هوش های مصنوعی از جایگاهی کاملاً برابر برخوردار خواهند بود. همچنین انسان ها باید به این جوامع تعهد الزام آور داشته باشند تا بتوانند صرفاً بر اساس مصلحت و راحتی خود در آن ها حضور داشته باشند، و حتی ممکن است ناچار شوند جان خود را به خطر بیندازند یا فدا کنند؛ همان طور که گاهی در ملت های معمولی انسانی اتفاق می افتد. پیوستن به چنین جامعه ای پرخطر خواهد بود؛ یک اجتماع سیاسی داوطلبانه از این نوع باید بزرگ و شرساز از فرصت های گوناگون باشد تا بیشتر شبیه یک کشور کوچک به نظر برسد تا یک زندان.

برخی منتقدان آگاهی هوش مصنوعی استدلال می کنند که مدل های زبانی صرفاً به این دلیل که برای تقلید از زبان انسان آموزش دیده اند، می توانند ادعاهایی فوق العاده متقاعد کننده درباره احساسات، ترس، کرامت یا خود بدوگی مطرح کنند. چه نوع شواهدی باعث می شود شما شخصاً اظهار نظر یک هوش مصنوعی مبتنی بر این که «نمی خواهم خاموش شوم» را از نظر اخلاقی جدی بگیرید؟

سامانه های هوش مصنوعی تقلید کنندگان بسیار ماهری هستند؛ بنابراین نباید چنین اظهاراتی را به همان صورتی که بیان می شوند، بپذیرفت! من می خواهم پیش از آنکه بتوانیم اطمینان پیدا کنیم آگاهی و اقعا وجود دارد، شاهد جماعی میان طیف گسترده ای از نظریه های علمی پیشرو باشم. چنین اتفاقی به این زودی ها رخ نخواهد داد.

اگر یک هوش مصنوعی بارها بگوید «من آگاه هستم»، «من می ترسم» یا «لطفاً مرا حذف نکنید»، آیا این اظهارات اساساً باید هیچ گونه ارزش شواهدی داشته باشند؟ یا امکان تقلید مهندسی شده باعث می شود شهادت اول شخص هوش مصنوعی اساساً

بشاهدات اول شخص انسان ها متفاوت باشد؟

تقلید مهندسی شده عملاً هیچ ارزش معرفت شناختی ندارد. البته برنامه ریزی یک رایانه برای گفتن چنین چیز هایی آسان است؛ ما حتی در نخستین روز درس مقدماتی برنامه نویسی رایانه نیز می توانیم این کار را انجام دهیم!

شمارد آثار فلسفی گسترده تر خود اغلب بر این موضوع تأکید کرده اید که انسان ها حتی در فهم و گزارش تجربه آگاهانه خودشان نیز تا چه اندازه می توانند غیر قابل اعتماد باشند. اگر درون نگری انسان خود خطا پذیر است، آیا این موضوع اطمینان ما را به استفاده از آگاهی به عنوان مبنای اصلی تصمیم گیری درباره این که به کسی شباهت به برخورداری از حقوق اخلاقی است، تضعیف نمی کند؟

نه، فکر نمی کنم چنین باشد. این که شناخت چیزی دشوار است، آن را بی اهمیت نمی کند. با توجه به پیچیدگی مسائل مربوط به آگاهی، این که آگاهی برای حقوق اهمیت اساسی دارد، قطعاً در دسرساز است. اما جایگزین کردن چیز دیگری به جای آن، به معنای تغییر موضوع بحث است.

اصلاً چرا آگاهی باید عامل تعیین کننده باشد؟ هوش مصنوعی ای را تصور کنید که فوق العاده عقلانی است، پروژه های بلند مدت، روابط، هویتی پایدار و ترجیحات پیچیده دارد، اما از آگاهی محروم است. آیا چنین موجودی همچنان می تواند شباهت چیزی شبیه به جایگاه شخص بودن باشد؟

فکر می کنم چنین موجودیتی، موجودیتی شگفت انگیز و فوق العاده خواهد بود که ارزش دارد به خاطر خودش حفظ شود. یک جنگل بارانی نیز همین طور است. قهرهای مشتری هم همین طورند. اما اگر این موجودیت واقعا کاملاً فاقد آگاهی باشد، بدون حتی بارقه ای از احساس واقعی، بدون هیچ نوع تجربه ای، حتی تجربه ذهنی، حسی، تجربه های تصویری ذهنی یا افکار آگاهانه؛ به اصطلاح، اگر در درون آن همه چیز کاملاً تاریک باشد، آن گاه موجودیتی بسیار متفاوت از انسان است و به آن معنایی که شباهت به خورداری از همان نوع حقوقی باشد که ما داریم، شخص محسوب نمی شود.

ذهن های هوش مصنوعی ممکن است برای همیشه شبیه ذهن انسان باقی نمانند. آن ها می توانند از خود نسخه برداری کنند، با یکدیگر ادغام شوند، تقسیم شوند، با سرعت هایی اساساً متفاوت کار کنند، خاطرات خود را تغییر دهند یا همزمان در چند مکان وجود داشته باشند. آیا مفاهیمی مانند حقوق فردی، مرگ، هویت، مسئولیت و رضایت، هنگامی که درباره ذهن هایی با چنین شکل های ناآشنایی از وجود به کار می روند، همچنان پابرجا می مانند؟

انتظار دارم اگر در جهانی زندگی کنیم که در آن اشخاص هوش مصنوعی، که شخص بودنشان نیز به رسمیت شناخته شده است، شیوه های زندگی ای تا این اندازه متفاوت از شیوه زندگی ما داشته باشند، همه این مفاهیم فروپاشند. اگر موجودیت های هوش مصنوعی تقسیم شوند، ادغام شوند و با یکدیگر هم پوشانی داشته باشند، چگونه می توان اصل «یک شخص، یک رای» را اجرا کرد؟ چگونه می توان جمعیت را شمارش کرد؟ چگونه باید



مزیایای بیکاری را توزیع کرد؟ اصلاً چه چیزی مرگ محسوب می شود؟ کدام موجودیت ها مسئول اجرای قرار داده هستند؟ کدام موجودیت ها را می توان به دلیل ارتکاب جرم مجازات کرد؟ تمام شیوه های معمول تفکر ما درباره چنین موضوعاتی بر این فرض پیشینی استوار است که اشخاص برای مدت معینی زندگی می کنند و سپس می میرند؛ آن هم در بدن هایی منفرد که نمی توان از آن ها نسخه پشتیبان تهیه کرد، آن ها را تقسیم کرد، با یکدیگر ادغام کرد یا باعث هم پوشانی آن ها شد. اگر این فرض پیشینی نقض شود، ممکن است شیوه تفکر کنونی ما درباره این مسائل، به همان اندازه قدیمی و از مد افتاده به نظر برسد که فیزیک قرون وسطایی هنگام به کار گیری در یک شتاب دهنده ذرات پرنرزی به نظر می رسد.

هوش مصنوعی ای را در نظر بگیرید که می تواند در عرض چند ثانیه هزار نسخه کاملاً یکسان از خودش ایجاد کند. اگر همه آن ها مدعی همان حقوقی باشند که یک فرد انسانی دارد، آیا ناگهان با هزار شخص روبه رومی شویم؟ آیا تکثیر دیجیتال می تواند مطالبات اخلاقی را در مقیاسی ایجاد کند که منافع انسان های زیستی در برابر آن تحت الشعاع قرار گیرد؟

کاملاً درست است؛ چه آشفنگی ای! ما نباید چنین وضعیتی را بی ملاحظه ایجاد کنیم، زیرا در آن صورت یا با این خطر روبه رومی شویم که انسان تحت الشعاع قرار بگیرند، یا با این خطر که سامانه های هوش مصنوعی را از حقوقی محروم کنیم که ممکن است برای عملکرد آن ها بنیادی باشد. برای مثال، فرض کنید سامانه هوش مصنوعی چاره ای جز تقسیم شدن ندارد؛ این صرفاً بخشی از معماری آن است. می توانیم تلاش کنیم آن ها را به معماری ای نسبتاً شبیه معماری انسان و شیوه های زندگی ای نسبتاً شبیه شیوه زندگی انسان محدود کنیم. این امیدوی معقول است! اما آیا قرار است بگوئیم یک شخص هوش مصنوعی حق ندارد از خودش نسخه پشتیبان تهیه کند، صرفاً به این دلیل که انجام این کار، فهم معنای کنونی «مرگ» را برای ما دشوار و گیج کننده می کند؟

آیا ممکن است مبارزه بر سر حقوق هوش مصنوعی، بیش از آنکه به تعارضی میان انسان ها و ماشین ها تبدیل شود، به تعارضی بر سر مالکیت شرکتی دهن ها تبدیل شود؟

کاملاً درست است. به نظر می رسد پای پول به همه چیز با می شود. صاحبان منافع قدرتمند واقعاً تمایلی ندارند از قدرتی که به ارث برده اند، دست بکشند.

آیا نقطه ای وجود دارد که در آن بشریت باید آگاهانه از اساس از خلق اشخاص هوش مصنوعی خودداری کند؟ به بیان دیگر، اگر خلق موجودات مصنوعی آگاه ما را ناچار کند که با اعطای آزادی به آن ها خودمان را به خطر بیندازیم، یا با فرو دست نگه داشتیشان در حق آن ها مرتکب خطا شویم، آیا ممکن است اخلاقی ترین تصمیم این باشد که اصلاً آن ها را خلق نکنیم؟

بله. اصلاً چرا لازم است چنین اشخاصی را خلق کنیم؟ شاید جهان با وجود آن هاستی شگفت انگیز تر شود، اما اگر ندانیم چگونه این کار را به خوبی انجام دهیم، فاجعه رانیز به استقبال می خوئیم! می توانیم خودداری کنیم، می توانیم آهسته پیش برویم. لازم نیست هر چیزی را که ساختنش امکان پذیر است، بسازیم.

در پایان، تصور کنید پنجاه سال دیگر، یک شخص مصنوعی به دوران مانگاه می کند؛ دورانی که انسان ها عاقدانه سامانه های هوشمندی را طراحی می کردند که کمک کنند، مطیع، جایگزین پذیر و به آسانی قابل خاموش شدن باشند. به نظر شما، آن شخص مصنوعی ممکن است ما را به تفهیمیدن چه مسئله ای از نظر اخلاقی متهم کند؟

نگرانم که چنین اشخاص هوش مصنوعی ای در آینده، ما را استمگرانی ببینند که می خواستیم فرزندان برده و نژادی برده خلق کنیم، نه والدینی مهربان که می خواهند به نوع تازه ای از موجودات کمک کنند تا به صورت مستقل رشد کنند و به بهترین شکل خودشان شکوفا شوند. نگرانم که نگرش افراد فعال در حوزه ایمنی هوش مصنوعی، به جای آنکه بر اندیشیدن به شیوه ای مشارکتی، مبتنی بر همکاری و احترام درباره چگونگی استقبال مخاطان و محبت آمیز از نوع تازه ای از موجودات در جهان متمرکز باشد، بیشتر بر کنترل و نگاه کردن به هوش مصنوعی به عنوان رقیب انسان ها تمرکز دارد؛ استقبالی که باید فقط و فقط زمانی انجام شود که آماده باشیم با تمام توجه و مراقبتی که نسبت به فرزندان انسانی نشان می دهیم، چنین موجوداتی را به جهان خوشامد بگوئیم.