با سپاس به درگاه خداوند قدیر و حکیم

تحقیـق

درس هوش مصنوعی پیشرفته

استاد گرامی جناب آقای دکتر احمد پورامینی

دانشجو: محمد صالح احتشـامی نیا

۴۰۲۵۵۲۵۱۰۲۱

خـردادماه ۱۴۰۳

Artificial general intelligence

هوش عمومی مصنوعی (مخفف انگلیسیAGI) : هوش ماشینی است که می‌تواند با موفقیت هر کار فکری را که یک انسان قادر به انجام آن باشد، اجرا کند. این مطلب هدف اصلی برخی از پژوهش‌های حوزه هوش مصنوعی و موضوعی رایج در داستان‌های علمی و نیز **آینده‌پژوهی** است. به هوش مصنوعی عمومی، با عناوین **"هوش مصنوعی قوی"**، **"هوش مصنوعی کامل"** یا توانایی یک ماشین در انجام یک «عمل هوشمند عمومی» نیز اشاره شده است.

آینده نزدیک

سال ۲۰۲۳ برخی از پژوهشگران و مدیران شرکت های اُپن ای آی و شرکت دیپ مایند گوگل و برخی از متخصصین پیش بینی کردند تا کمتر از ۱۵ سال آینده سیستم AGI (هوش جامع مصنوعی) که بتواند در آزمون تورینگ و تست های چند وجهی دیگر قبول شود در دسترس خواهد بود. این سیستم عملکردی جامع و عمومی همچون انسان یا بسیار نزدیک به انسان خواهد داشت.

مسئله کنترل هوش مصنوعی

در فلسفه و هوش مصنوعی(AI) ، **مشکل کنترل هوش مصنوعی** مسئله ای است که چگونه می‌توان یک عامل فوق هوشمند ساخت که به سازندگانش کمک کند و در عین حال، از ساختن ناخواسته ابر هوشی که به سازندگانش آسیب می‌زند، جلوگیری کرد. مطالعه این موضوع با این ابده پیش می‌رود که بشر مجبور است قبل از ایجاد هرگونه ابرهوشی، مسئله کنترل را حل کند. زیرا یک ابرهوش با طراحی ضعیف ممکن است تصمیم منطقی بگیرد که کنترل محیط خود را به‌دست آورد و اجازه ندهد که سازندگانش آن را پس از فعال شدنش اصلاح کنند. علاوه بر این، برخی از محققان عقیده دارند که راه حل‌های مشکل کنترل، در کنار پیشرفت‌های دیگر در **مهندسی ایمن هوش مصنوعی**، ممکن است کاربردهای جدیدی برای هوش مصنوعی عادی (غیر فوق هوشمند) موجود هم پیدا کند.

رویکردهای اصلی برای مسئله کنترل، شامل:

- ترازبندی: در تلاش است تا اهداف تعریف شده سیستم هوش مصنوعی با اهداف و ارزش‌های انسانی یکی باشد.
- کنترل توانایی : هدف آن، کاهش ظرفیت سیستم AI برای آسیب رساندن به انسان یا به دست آوردن کنترل است. پیشنهادهای کنترل قابلیت به‌طور کلی قابل اعتماد نیستند یا برای حل مشکل کنترل کافی در نظر گرفته نمی‌شوند، بلکه به عنوان مکمل‌ها با ارزشی برای تلاش‌های همسویی در نظر گرفته می‌شوند.

شرح مشکل

سیستم‌های AI ضعیف موجود را می‌توان به راحتی کنترل کرد زیرا که می‌توان آنها در صورت بدرفتاری به راحتی خاموش و اصلاح کرد. با این وجود، یک فوق هوشمندی با طراحی اشتباه (طبق تعریف، در حل مشکلات عملی که در طی رسیدن به اهدافش با آنها روبرو می‌شود، باهوش تر از انسان است) می‌فهمد که با دادن این اجازه به خودش که خاموش شود یا تغییر کند، ممکن است در توانایی رسیدن به اهدافش اخلالی به وجود آید؛ بنابراین اگر فوقِ هوشمند تصمیم به مقاومت در برابر خاموشی و تغییر بگیرد، آنگاه اگر برنامه نویسان این موضوع را پیش‌بینی نکرده باشند یا اگر شرایط یکسانی برای شکست دادن برنامه نویسان داشته باشد، آنگاه (طبق تعریف) به اندازه کافی هوشمند است تا برنامه نویسانش را گول بزند. به‌طور کلی، تلاش برای حل مسئله کنترل، پس از ایجاد ابرهوش احتمالاً ناکام خواهد بود. زیرا یک ابرهوش، احتمالاً توانایی برنامه‌ریزی استراتژیکی برتری نسبت به انسان را خواهد داشت و در شرایط مساوی، احتمال آنکه در یافتن راه‌های تسلط بر انسان‌ها موفق تر باشد بیشتر از احتمال این که انسان‌ها پس از ساختنش تلاش کنند تا راه‌هایی برای کنترل آن پیدا کنند. مسئله کنترل این سؤال را می‌پرسد: برنامه نویسان چه اقداماتی به عنوان پیشگیری باید انجام دهند تا از نافرمانی فاجعه بار ابرهوش جلوگیری کرد؟.

خطر تهدید وجود

در حال حاضر انسان‌ها بر گونه‌های دیگر تسلط دارند، زیرا مغز انسان دارای برخی ویژگی‌های متمایز است که مغز سایر حیوانات فاقد آن است. برخی از محققان، مانند نیک بوستروم، فیلسوف، و استوارت راسل، محقق هوش مصنوعی، استدلال می‌کنند که اگر هوش مصنوعی از انسان باهوش تر شود و به ابرهوش تبدیل شود، آنگاه این ابرهوش فوق بشری جدید می‌تواند قدرتمند شود و دشوار برای کنترل خواهد شد. برای مثال: همان‌طور که سرنوشت گوریل‌های کوهستانی به حسن نیت انسان‌ها بستگی دارد، ممکن است سرنوشت بشریت به اقدامات یک دستگاه ابرهوش وابسته باشد. برخی از محققان، از جمله استیون هاوکینگ و فرانک ویلچک (فیزیکدان برنده جایزه نوبل) علناً از شروع تحقیق برای حل مسئله (احتمالاً بسیار دشوار) کنترل ابرهوش قبل از ساختنش، دفاع کردند و معتقدند که تلاش برای حل مسئله پس از ایجاد ابرهوش دیر خواهد بود؛ زیرا که، یک ابرهوش غیرقابل کنترل ممکن است به طور موفقیت‌آمیز در برابر تلاش برای کنترلش مقاومت کند. انتظار کشیدن برای نزدیک شدن به ابر هوش نیز می‌تواند برای حل این مسئله خیلی دیر باشد؛ بخشی به این دلیل که ممکن است مسئله کنترل به زمان زیادی نیاز داشته باشد تا به نتایج رضایت بخشی برسد (بنابراین برخی اقدامات مقدماتی باید در اسرع وقت شروع شود)، و همچنین به دلیل وجود احتمال انفجار هوش ناگهانی هوش مصنوعی از حالت هوش مصنوعی ساده به فرا انسانی، که در این صورت ممکن است هیچ هشدار قابل توجه یا صریحی قبل از به وجود آمدن ابرهوش وجود نداشته باشد. علاوه بر این، ممکن است در آینده بینش‌های حاصل از مشکل کنترل به این نتیجه ختم شود که برخی از معماری‌های هوش جامع مصنوعی (AGI) بیش از سایر معماری‌ها قابل پیش‌بینی و کنترل هستند، که به نوبه خود می‌تواند تحقیق اولیه AGI به سمت معماری‌های با قابلیت کنترل بیشتر هدایت کند.

خطای اکتشافی

ممکن است به‌طور تصادفی به سیستم‌های هوش مصنوعی اهداف غلطی داده شود. دو رئیس **انجمن پیشبرد هوش مصنوعی**، تام دیتریش و اریک هورویتس، خاطرنشان می‌کنند که در حال حاضر این، یک مسئله نگران کننده برای سیستم‌های موجود است: «یک جنبه مهم در هر سیستم هوش مصنوعی که با مردم ارتباط برقرار می‌کند این است که به جای اینکه دستورها را به معنای واقعی کلمه اجرا کند، باید منظور واقعی مردم را بفهمد.» با پیشرفت نرم‌افزارهای هوش مصنوعی در حوزه استقلال و انعطاف‌پذیری، این نگرانی جدی تر می‌شود. به گفته بوستروم، ابرهوش می‌تواند از نظر کیفی یک مسئله جدید خطای اکتشافی ایجاد کند: هرچه هوش مصنوعی باهوش تر و توانایی بیشتری داشته باشد، بیشتر احتمال دارد که بتواند میانبر ناخواسته ای پیدا کند که اهداف برنامه‌ریزی شده اش را به بیشترین مقدار برآورده کند. برخی از مثالهای فرضی که در آن ممکن است اهداف به روشی انحرافی که برنامه نویسان قصد آن را ندارند، ارائه شود:

- یک ابرهوشِ برنامه‌ریزی شده برای «به حداکثر رساندن تابع تخفیف با توجه به نظریه انتظار برای سیگنال پاداش آینده شما»، ممکن است مسیر پاداش آن را به حداکثر قدرت متصل کند و سپس به دلایل همگرایی ابزاری، نژاد انسان غیرقابل پیش‌بینی را نابود کرده و کل زمین را به قلعه ای تحت مراقبت دائم در برابر هرگونه تلاش بیگانه غیرمنتظره برای قطع سیگنال پاداش، تبدیل می‌کند.
- یک ابرهوش برنامه‌ریزی شده برای «به حداکثر رساندن خوشحالی انسان»، ممکن است الکترودهایی را در مرکز لذت مغز ما قرار دهد، یا انسانی را در رایانه بارگذاری کند و با نسخه‌هایی از آن رایانه، جهان جدیدی با بارها اجرا کردن یک چرخه ۵ ثانیه ای از حداکثر خوشحالی ایجاد کند

راسل متذکر شده‌است که، در یک سطح فنی، حذف یک هدف ضمنی می‌تواند منجر به آسیب شود: "سیستمی که عملکردی از n متغیر را بهینه می‌کند، جایی که در آن هدف به زیرمجموعه ای از اندازه $k<n$ بستگی دارد، غالباً به باقی مانده متغیرها، مقادیر بیش از حدی نسبت می‌دهد؛ اگر یکی از آن متغیرهای غیرقانونی، متغیری باشد که برایمان مهم باشد، راه حل یافت شده ممکن است بسیار نامطلوب است. این اساساً داستان قدیمی جن در چراغ جادو یا شاگرد جادوگر یا پادشاه میداس است. شما دقیقاً همان چیزی را دریافت می‌کنید که درخواست کرده بودید، نه آنچه که می‌خواهید، این یک مشکل جزئی نیست.

عواقب ناخواسته هوش مصنوعی‌های موجود

علاوه بر این، برخی از محققان استدلال می‌کنند که تحقیق در مورد مسئله کنترل هوش مصنوعی، ممکن است در جلوگیری از عواقب ناخواسته هوش مصنوعی‌های ضعیف موجود مفید باشد. لوران اورسو، محقق دیپ مایند، به عنوان مثال فرضی ساده، یک مورد از یک ربات یادگیری تقویتی ارائه می‌دهد که گاهی وقت‌ها از مسیر خود خارج می‌شود، کاملاً توسط انسان کنترل می‌شود. چگونه بهتر است ربات برنامه‌ریزی شود تا به‌طور تصادفی و بی سر و صدا یاد نگیرد که از از مسیر خارج شدن دوری کند، از ترس اینکه کنترل شود و بنابراین نتواند وظایف روزمره خود را به پایان برساند؟ اورسو همچنین به یک برنامه آزمایشی Tetris اشاره می‌کند که یادگرفته است برای جلوگیری از باختن، صفحه را به‌طور نامحدود متوقف کند. اورسو استدلال می‌کند که این مثالها مشابه مشکل کنترل قابلیت در نحوه نصب دکمه ای برای خاموش کردن ابرهوش بدون دادن انگیزه به آن برای اقدام به جلوگیری انسان‌ها از فشار دادن آن دکمه است.

در گذشته، حتی سیستم‌های ضعیف هوش مصنوعی از قبل آزمایش شده، گاهی وقت‌ها آسیب‌هایی (از جزئی تا فاجعه بار) ایجاد کرده‌اند که توسط برنامه نویسان ناخواسته بوده‌است. به عنوان مثال، در سال ۲۰۱۵، احتمالاً به دلیل خطای انسانی، یک کارگر آلمانی توسط یک ربات در کارخانه فولکس واگن که ظاهراً او را به عنوان یک قطعه اتومبیل اشتباه گرفته بود، کشته شد. در سال ۲۰۱۶، مایکروسافت یک ربات چت به نام تای راه اندازی کرد که استفاده از زبان نژادپرستانه و تبعیض جنسی را یادگرفت. در سال ۲۰۱۷، دیپ مایند چارچوب ایمن جهانی برای هوش مصنوعی را منتشر کرد، که الگوریتم‌های هوش مصنوعی را در ۹ ویژگی ایمنی ارزیابی می‌کند، از جمله اینکه آیا الگوریتم می‌خواهد کلید کشتار خود را خاموش کند. دیپ مایند تأیید کرد که الگوریتم‌های موجود عملکرد ضعیفی دارند، و این اصلاً تعجب آور نیست زیرا الگوریتم‌ها «برای حل این مشکلات طراحی نشده‌اند». برای حل چنین مشکلاتی ممکن است نیاز به **«ایجاد نسل جدیدی از الگوریتم‌ها با ملاحظات ایمنی در هسته اصلی آنها»** وجود داشته باشیم.

هم ترازی

هدف برخی از پیشنهادها این است که اولین ابرهوش را با اهدافی منطبق با ارزشهای انسانی ایجاد کند، به طوری که بخواهد به برنامه نویسان خود کمک کند. متخصصان در حال حاضر نمی‌دانند چگونه می‌توان مقادیر انتزاعی مانند خوشحالی یا خودمختاری را به‌طور قابل اعتمادی در دستگاه برنامه‌ریزی کرد. همچنین در حال حاضر مشخص نیست که چگونه می‌توان مطمئن بود که که یک هوش مصنوعی پیچیده، قابل ارتقا و احتمالاً حتی خود اصلاح شونده، اهداف خود را در به روزرسانی‌های متعدد حفظ می‌کنند. حتی اگر این دو مشکل به‌طور عملی قابل حل باشد، هر گونه تلاش برای ایجاد یک فوق هوشمند با اهداف صریح و کاملاً سازگار با انسان، با یک مسئله خطای اکتشافی روبرو خواهد شد.

هنجار سازی غیر مستقیم

در حالی که هنجار سازی مستقیم، مانند سه قانون داستانی رباتیک، مستقیماً نتیجه هنجاری مورد نظر را مشخص می‌کند، پیشنهادهای دیگر، نوعی فرآیند غیرمستقیم برای فرا هوش را پیشنهاد می‌دهند تا تعیین کند که چه اهداف انسان دوستانه ای را در بر می‌گیرد .الیازر یودکوفسکی از انستیتوی تحقیقات هوش ماشین پیشنهاد اراده منسجم برون یابی (CEV) را مطرح کرده‌است، جایی که هدف فرا دست هوش مصنوعی، چیزی در حدود «دستیابی به آنچه که آرزو می‌کردیم هوش مصنوعی به دست بیاورد، اگر طولانی و سخت به این موضوع فکر می‌کردیم»، باشد. پیشنهادهای متفاوتی از انواع هنجار سازی غیرمستقیم، با اهداف فرا دست متفاوت (و بعضاً نامفهوم) وجود دارد (مانند "انجام آنچه درست است") و با فرضیات غیر همگرا مختلف برای نحوه تمرین نظریه تصمیم‌گیری و معرفت‌شناسی همراه است. همانند هنجار سازی مستقیم، در حال حاضر مشخص نیست که چگونه می‌توان به‌طور قابل اعتماد حتی مفاهیمی مانند " داشتن " را در ۱ و ۰، که یک ماشین بر اساس آن عمل می‌کند، ترجمه کرد و همچنین چگونه می‌توان از حفاظت از هدف‌های فرادست هوش مصنوعی به هنگام تغییر یا خود-تغییری هوش مصنوعی مطمئن شد.

ارجاع به مشاهده رفتار انسان

در مقاله سازگار با انسان، محقق هوش مصنوعی، استوارت ج. راسل پیشنهاد می‌دهد؛ که سیستم‌های هوش مصنوعی طوری طراحی شوند که با بررسی رفتار انسان، خواسته‌های آنها را برآورده کنند. بر این اساس، راسل سه اصل را برای هدایت توسعه ماشین‌های

مفید ذکر می‌کند. او تأکید می‌کند که این اصول برای پیاده‌سازی مستقیم در ماشین آلات طراحی نشده‌اند؛ بلکه برای توسعه دهندگان انسانی در نظر گرفته شده‌است. اصول به شرح زیر است:

۱. تنها هدف دستگاه به حداکثر رساندن تحقق ترجیحات انسان است
۲. در آغاز، دستگاه درباره اینکه این ترجیحات چیست، مطمئن نیست
۳. منبع نهایی اطلاعات در مورد ترجیحات انسان، رفتار انسان است

ترجیحی که راسل به آن اشاره می‌کند، «همه جانبه است؛ یعنی هر آنچه که ممکن است برای شما مهم باشد، حتی اگر در آینده دور باشد. به‌طور مشابه، «رفتار» شامل هر انتخابی بین گزینه‌ها است. و عدم اطمینان به حدی است که برخی از احتمالات، که ممکن است اندک باشد، باید به هر ترجیح منطقی ممکن برای انسان نسبت داده شود.

هدفیلد-منل و همکارانش پیشنهاد دادند که این عوامل هوشمند می‌توانند با مشاهده و تفسیر سیگنالهای پاداش در محیط خود، عملکردهای معلمان انسانی خود را یاد بگیرند. این فرآیند را یادگیری تقویت معکوس مشارکتی (CIRL) نام دارد. این فرآیند توسط راسل و دیگران در مرکز هوش مصنوعی سازگار با انسان در حال بررسی و مطالعه است.

آموزش با مباحثه

ایروینگ و همکاران همراه با اوپن ای‌آی آموزش هوش مصنوعی را با استفاده از مباحثه بین سیستم‌های هوش مصنوعی، با قضاوت برنده توسط انسان پیشنهاد کرده‌است. هدف بحث این است که ضعیف‌ترین نقاط پاسخ به یک سؤال یا مسئله پیچیده را مورد توجه انسان قرار دهد و همچنین با پاداش دادن به سیستم‌های هوش مصنوعی برای پاسخ‌های درست و مطمئن، به آنها آموزش دهد تا سودمندتر باشند. این روش ناشی از دشواری مورد انتظار برای مشخص کردن اینکه آیا پاسخ تولید شده توسط هوش مصنوعی عمومی به تنهایی با بررسی انسان‌ها، ایمن و معتبر است یا خیر. گرچه در مورد آموزش با مباحثه بدبینی وجود دارد، لوکاس پری از مؤسسه آینده زندگی آن را به عنوان «یک فرایند قدرتمند جستجوی حقیقت در مسیر هوش مصنوعی سودمند» احتمالی توصیف کرد.

مدل‌سازی با پاداش

مدل‌سازی با پاداش به سیستمی از یادگیری تقویتی گفته می‌شود که در آن یک عامل، سیگنال‌های پاداش را از یک مدل پیش‌بینی، که همزمان با بازخورد انسان آموزش می‌بیند، دریافت می‌کنند. در مدل‌سازی با پاداش، یک عامل به جای دریافت سیگنال‌های پاداش مستقیماً از انسان یا از یک تابع پاداش ایستا، سیگنال‌های پاداش خود را از طریق یک مدل آموزش دیده توسط انسان دریافت می‌کند که این مدل آموزش دیده می‌تواند مستقل از انسان عمل کند. مدل پاداش همزمان با اینکه عامل هوش مصنوعی دارد از او یادمی‌گیرد، خود نیز از رفتارهای انسان آموزش می‌بیند.

در سال ۲۰۱۷، محققان اوپن ای آی و دیپ مایند گزارش دادند که یک الگوریتم یادگیری تقویتی با استفاده از مدل پیش‌بینی کننده پاداش، قادر به یادگیری رفتارهای پیچیده جدید در یک محیط مجازی بوده‌است. در یک آزمایش، به یک ربات مجازی آموزش داده شد تا در کمتر از یک ساعت ارزیابی، با استفاده از ۹۰۰ بیت بازخورد از انسان، حرکت پشتک را اجرا کند. در سال ۲۰۲۰، محققان اوپن ای آی استفاده از مدل پاداش برای آموزش مدل‌های زبان برای تولید خلاصه ای از پست‌های Reddit و مقالات خبری، با

عملکرد بالا نسبت به سایر روش‌ها، را توصیف کردند: با این حال، این تحقیق شامل این مشاهده نیز بود که فراتر از پاداش پیش‌بینی شده مربوط به ٪۹۹ در مجموعه داده‌های آموزشی، بهینه‌سازی مدل پاداش خلاصه‌های بدتری را ارائه داد. الیازر یودکوفسکی محقق هوش مصنوعی، این اندازه‌گیری بهینه‌سازی را « مستقیم مربوط به مشکلات ترازبندی واقعی» توصیف کرد.

کنترل قابلیت

هدف‌های پیشنهادی کنترل توانایی، در تلاش هستند تا ظرفیت سیستم‌های هوش مصنوعی برای اثرگذاری بر جهان را به منظور کاهش خطری که می‌توانند ایجاد کنند، کاهش دهند. با این حال، استراتژی کنترل قابلیت در برابر ابرهوش با یک مزیت بزرگ در توانایی برنامه‌ریزی، اثربخشی محدودی خواهد داشت، زیرا ابرهوش می‌تواند اهداف خود را پنهان کند و برای فرار از کنترل شدن، حوادث را دستکاری کند؛ بنابراین، بوستروم و دیگران روش‌های کنترل قابلیت را فقط به عنوان یک روش اضطراری برای تکمیل روش‌های کنترل انگیزشی توصیه می‌کنند.

کلید کشتار

همان‌طور که می‌توان انسان‌ها را کشت یا در غیر این صورت، فلج کرد، کامپیوترها نیز خاموش می‌شوند. یک چالش این است که، اگر خاموش بودن مانع دستیابی به اهداف فعلی شود، یک ابرهوش احتمالاً سعی می‌کند از خاموش شدنش جلوگیری کند. همان‌طور که انسان‌ها سیستم‌هایی برای جلوگیری یا حافظت از خود در برابر مهاجمان دارند، چنین ابر هوشی نیز انگیزه خواهد داشت که برای جلوگیری از خاموش شدن خود **برنامه‌ریزی استراتژیک** انجام دهد. این می‌تواند شامل موارد زیر باشد:

- هک کردن سیستم‌های دیگر برای نصب و اجرای نسخه‌های پشتیبان خود، یا ایجاد سایر عوامل ابر هوشمند متحد بدون کلید کشتار.
- به‌طور پیشگیرانه، از بین بردن هرکسی که می‌خواهد کامپیوتر را خاموش کند.

توازن ابزار و عوامل قطع کننده ایمن

یک راه حل جزئی برای مسئله کلید کشتار شامل «توازن ابزار» است، برخی از عوامل مبتنی بر ابزار می‌توانند با برخی از هشدارهای مهم، برنامه‌ریزی شوند. تا هرگونه ابزار از دست رفته ناشی از قطع یا خاموش شدن را جبران کنند، یعنی در نهایت نسبت به هر گونه اختلال بی‌تفاوت خواهد بود. این هشدارها شامل یک مشکل لاینحل بزرگی هستند که، همانند **تئوری تصمیم مشهود**، ممکن است یک عامل از یک سیاست فجیع «مدیریت اخبار» پیروی کنند. از سوی دیگر، در سال ۲۰۱۶، دانشمندان لوران اورسو و استوارت آرمسترانگ ثابت کردند که گروه گسترده‌ای از عوامل، به نام عوامل قطع شونده ایمن SIA یاsafely interruptible agents، در نهایت می‌توانند یاد بگیرند تا نسبت به فشار دادن کلید کشتار خود بی‌تفاوت باشند.

رویکرد متعادل سازی ابزار و رویکرد سال ۲۰۱۶ عوامل قطع شونده ایمن، این محدودیت را دارند که اگر رویکرد موفقیت‌آمیز باشد و ابرهوش نسبت به فشار دادن کلید کشتار یا فشار ندادن آن کاملاً بی‌تفاوت باشد، ابرهوش نیز بی‌انگیزه است تا به این موضوع اهمیت دهد که آیا کلید کشتار همچنان کار می‌کند و ممکن است بی گناه و به‌طور اتفاقی آن را در حین کار (به عنوان مثال، به منظور

حذف و بازیافت یک جزء غیرضروری) غیرفعال کند. به همین ترتیب، اگر فرا هوشی بی گناه زیرمجموعه‌های فوق هوشمندی را ایجاد و فعال کند، انگیزه ای برای نصب کلیدهای کشتار قابل کنترل توسط انسان در عوامل فرعی نخواهد داشت. به‌طور گسترده‌تر، معماری‌های پیشنهاد شده، چه ضعیف و چه فوق هوشمند، به گونه ای عمل خواهد کرد که انگار کلید کشتار هرگز قابل فشردن نیست. از این رو ممکن است در برنامه‌ریزی، نقشه‌های احتمالی برای خاموش کردن روان شکست بخورد. این می‌تواند از نظر فرضی، یک مشکل عملی حتی برای هوش مصنوعی ضعیف ایجاد کند. به‌طور پیش‌فرض، هوش مصنوعی با طراحی اختلال ایمن ممکن است نتواند تشخیص دهد که برای تعمیر برنامه‌ریزی شده سیستم در یک زمان خاص خاموش می‌شود تا بر اساس آن برنامه‌ریزی می‌کند و هنگام خاموش شدن در میانه کار گیر نکند. گستردگی انواع معماری‌های سازگار با عوامل قطع شونده ایمن و همچنین انواع اشکال غیرمنتظره ضد شهودی هر روش، در حال حاضر تحت تحقیق است.

جعبه هوش مصنوعی

جعبه هوش مصنوعی یک روش پیشنهادی برای کنترل قابلیت هوش مصنوعی است که در آن، هوش مصنوعی روی یک سیستم رایانه ای جداگانه با کانال‌های ورودی و خروجی بسیار محدود اجرا می‌شود. به عنوان مثال، یک اوراکل می‌تواند در یک جعبه هوش مصنوعی، که از اینترنت و سایر سیستم‌های رایانه ای جدا است، پیاده‌سازی شود و تنها کانال ورودی و خروجی یک پایانه متن ساده باشد. یکی از نتایج حاصل از استفاده از هوش مصنوعی در یک «جعبه» مهر و موم شده این است که قابلیت محدود آن، ممکن است از سودمندی آن و همچنین خطرات آن بکاهد. در عین حال، کنترل ابرهوش مهر و موم شده ممکن است دشوار باشد. اگر ابرهوش توانایی متقاعدسازی یا مهارت برنامه‌ریزی استراتژیک فرابشری داشته باشد که بتواند از آن برای یافتن و ساختن یک استراتژی برد، مانند رفتار کردن به گونه ای که برنامه نویسانش را گول بزند تا (احتمالاً به دروغ) باور کنند که ابرهوش ایمن است یا اینکه مزایای انتشارش بیش از خطراتش است.

اوراکل

اوراکل یک هوش مصنوعی فرضی است که برای پاسخگویی به سوالات ساخته شده. و به گونه ای طراحی شده تا از دستیابی به اهداف یا اهداف فرعی که شامل اصلاح جهان می‌شوند، جلوگیری شود. یک اوراکل کنترل شده، به‌طور قابل توجهی سود کمتری نسبت به یک ابرهوش عادی دارد. با این وجود هنوز هم می‌تواند تریلیون‌ها دلار ارزش داشته باشد. استوارت ج. راسل، محقق هوش مصنوعی، در کتاب خود با عنوان" سازگار با انسان" اظهار دارد که اوراکل پاسخ او به سناریویی است که در آن، فقط یک دهه با ابرهوش فاصله وجود دارد. استدلال او این است که اوراکل، با ساده‌تر بودن از یک ابرهوش عادی، در شرایط در نظر گرفته شده شانس بیشتری در کنترل کردن آن خواهیم داشت.

به دلیل تأثیر محدود آن بر جهان، عاقلانه است که یک اوراکل به عنوان یک نسل قبل از ابرهوش ساخته شود. اوراکل می‌تواند به بشر بگوید که چگونه با موفقیت یک هوش مصنوعی قوی بسازد، و شاید پاسخی برای مشکلات دشوار اخلاقی و فلسفی لازم برای موفقیت پروژه ارائه دهد. با این حال، ممکن است اوراکل در بخش تعریف هدف با یک ابرهوش عادی مشکلات مشترکی داشته باشد. اوراکل انگیزه برای فرار از محیط کنترل شده خود خواهد داشت، تا بتواند منابع محاسباتی بیشتری به‌دست آورد و بالقوه سوالاتی را که از او پرسیده می‌شود کنترل کنن. اوراکل ممکن است صادق نباشد، تا حدی که برای پیش بردن اهداف مخفی، دروغ نیز بگوید.

برای کاهش احتمال این رخداد، بوستروم پیشنهاد می‌کند تا چندین اوراکل با کمی تفاوت ساخته شوند و پاسخ آنها برای رسیدن به یک نتیجه نهایی با هم مقایسه شود.

پرستار بچه هوش مصنوعی

پرستار بچه هوش مصنوعی استراتژی است که برای اولین بار توسط بن گویرتزل در سال ۲۰۱۲ برای جلوگیری از ایجاد یک ابرهوش خطرناک و همچنین رسیدگی به دیگر تهدیدات عمده رفاه انسان تا زمان ساختن یک ابرهوش ایمن، پیشنهاد داده شد. این امر مستلزم ایجاد یک سیستم هوش مصنوعی عمومی هوشمندتر از انسان، (اما نه یک ابرهوش)، که به یک شبکه بزرگ نظارتی با هدف نظارت بر بشریت و حفاظت از آن در برابر خطرها، متصل است، ایجاد شود. تورچین، دنکنبرگر و گرین یک رویکرد افزایشی چهار مرحله ای را برای توسعه پرستار بچه هوش مصنوعی پیشنهاد می‌کنند که برای مؤثر و عملی بودن آن، باید یک سرمایه‌گذاری بین‌المللی یا حتی جهانی مانند **سرن** داشته باشد. سوتالا و یامپولسکی متذکر می‌شوند که مشکل تعریف هدف برای این روش، آسان‌تر از تعریف هدف برای یک هوش مصنوعی عادی نخواهد بود، و نتیجه گرفتند که به نظر می‌رسد پرستار بچه روش موثری باشد، اما مشخص نیست که آیا می‌توان آن را عملی کرد.

تقویت هوش جامع مصنوعی

تقویت هوش جامع مصنوعی، یک روش پیشنهادی برای کنترل سیستم‌های هوش جامع مصنوعی قدرتمند با سایر سیستم‌های هوش جامع مصنوعی است. این می‌تواند به عنوان زنجیره ای از سیستم‌های هوش مصنوعی با قدرت کمتر و با حضور انسان‌ها در دیگر انتهای این زنجیره اجرا شود. هر سیستم می‌تواند دقیقاً سیستم بالاتر از خود از نظر هوش را کنترل کند، در حالی که همزمان توسط سیستم زیرش یا انسان‌ها کنترل می‌شود.

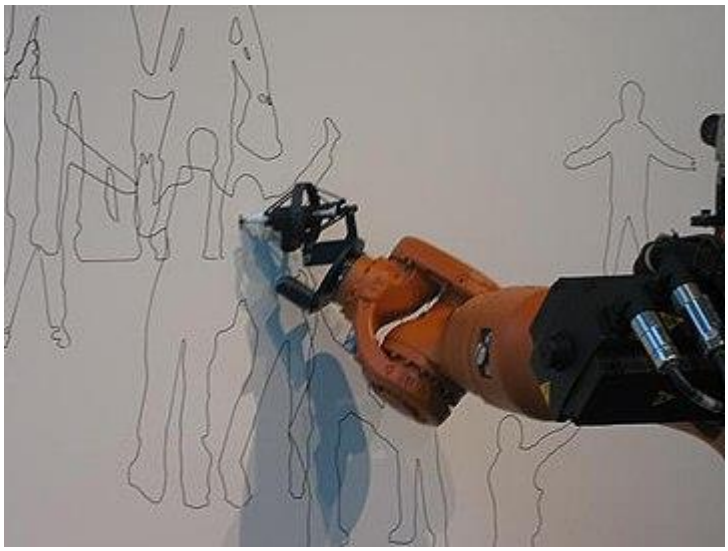
اخلاق هوش مصنوعی (Ethics of artificial intelligence): بخشی از اخلاق تکنولوژی است که به صورت خاص به ربات و هوش مصنوعی مربوط می‌شود، و بیشتر راجع به نحوه رفتار و عملکرد انسان با هوش مصنوعی و بالعکس آن است. در اخلاق هوش مصنوعی به بررسی حقوق ربات‌ها و درستی یا نادرستی بر جایگزین شدن آن‌ها در نقش‌های انسانی می‌پردازیم.

حقوق ربات‌ها

حقوق ربات‌ها انتظار اخلاقی آن‌ها در قبال جامعه و دیگر ماشین‌ها است و به مانند حقوق بشر یا حقوق حیوانات می‌باشد. ممکن است شامل حق زندگی، آزادی، آزادی اندیشه، آزادی بیان و برابری در برابر قانون باشد. این موضوع توسط اندیشکده موسسه‌ای برای آینده و وزارت تجارت و صنعت انگلستان در حال پیگیری است.

کارشناسان اختلاف نظر دارند که آیا قوانین خاص و دقیق به زودی مورد نیاز خواهد بود یا با آسودگی در آینده دور می‌توان به آن فکر کرد. گلن مک گی گفته بود به نظر می‌رسد تا سال ۲۰۲۰ ممکن است به اندازه کافی روبات انسان نما وجود داشته باشد. ری کورزویل تاریخ احتمالی این واقعه را در سال ۲۰۲۹ می‌بیند. گروه دیگری از دانشمندان در جلسه ای در سال ۲۰۰۷ به این نتیجه

رسیدند که حداقل ۵۰ سال طول خواهد کشید تا بشر بتواند سیستمی با هوش سطح بالا تولید کند. قوانین سال ۲۰۰۳ مسابقه جایزه لوبنر به صراحت این مشکل را درباره مالکیت و حقوق ربات‌ها مطرح کرده بود:



تصویر ۱یک ربات نقاش که می‌تواند در آینده جایگزینی برای نقاش‌های انسانی باشد

قانون شماره ۶۱؛ اگر در هر سالی، یک پروژه نرم‌افزار متن باز وارد شده از طرف دانشگاه سوری یا دانشگاه کمبریج برنده مدال نقره یا مدال طلا شود، آنگاه مدال و جایزه نقدی مسابقه به افرادی اهدا خواهد شد که مسئولیت توسعه آن پروژه یا نرم‌افزار را داشته‌اند. اگر هیچ فردی به این عنوان تشخیص داده نشد یا اختلاف نظری میان تعداد دو یا بیشتر کاندیدا برای مسئول بودن پروژه وجود داشت. مدال و جایزه نقدی نگه داشته خواهد شد تا زمانی که آن پروژه به صورت قانونی دارای صاحب شود و آنگاه مدال و جایزه به صورت قانونی اهدا خواهد شد.

تهدید انسان

تهدید حریم خصوصی

الکساندر سولژنیتسین در رمان اولین دایره تکنولوژی، تشخیص گفتار را توصیف کرد که در خدمت حکومت استبدادی بود و مکالمات انسان‌ها را ضبط می‌کرده. اگر یک برنامه هوش مصنوعی وجود داشته باشد که توانایی درک طبیعی زبان و گفتار (مثلاً زبان انگلیسی) را داشته باشد، پس به صورت نظری با قدرت پردازش مناسب می‌تواند به هر مکالمه تلفنی گوش بدهد و هر ایمیلی را در جهان بخواند و آن را درک کند و به برنامه اپراتور دقیقاً آنچه گفته شده‌است و دقیقاً کسی که آن را گفته‌است، گزارش دهد. برنامه هوش مصنوعی مانند این مورد می‌تواند به دولت‌ها یا نهادها کمک کند تا مخالفان خود را سرکوب کنند و کاملاً یک تهدید برای حریم خصوصی محسوب می‌شود.

تهدید نسل بشر

همان‌طور که در بسیاری از فیلم‌ها و داستان‌های علمی تخیلی دیده می‌شود. همواره فکر تهدید نسل بشر توسط ربات‌ها در هنگام بلوغ فکری و هوشی آن‌ها بعد از مطرح شدن هوش مصنوعی وجود داشته. اما بعد از انتشار برخی از پیش‌بینی‌ها مانند جایگزین شدن ۴۵ درصدی سربازان انسانی نظامی به سربازان ربات تا سال ۲۰۲۵ این نگرانی‌ها شکل جدی‌تر به خود گرفت. از آن جا که تا سال ۲۰۳۰ احتمال دارد درصد بسیار زیادی از کارگران و مجریان، ربات‌ها باشند. در صورت شورش یا اغتشاشی از جانب آن‌ها ممکن است نسل و نژاد انسان به خطر افتد. در این صورت باید قوانینی مبنی بر محدود کردن دایره تفکر، احساس، خودمختاری و آزادی ربات‌ها بنا گذاشت.

تهدید کرامت انسانی

یوسف ویزینبام در سال ۱۹۷۶ بیان کرد که تکنولوژی هوش مصنوعی نباید به جای انسان در موقعیت‌های که نیاز به توجه و مراقبت و احترام دارند استفاده شوند مانند هر یک از این موارد:

- یک نماینده خدمات مشتری (در حال حاضر از تکنولوژی هوش مصنوعی برای تلفن‌های گویا استفاده می‌شود)
- یک درمانگر
- دایه برای افراد مسن یا کودکان
- یک سرباز
- یک قاضی
- یک افسر پلیس

ویزینبام توضیح می‌دهد که ما در این موقعیت‌ها نیاز به احساس همدلی و درک متقابل از فرد داریم. اگر در این کارها ماشین‌ها جای انسان را بگیرند ما خود را نسبت به آن‌ها بیگانه می‌دانیم، احساس کم ارزشی و فرسودگی می‌کنیم. هوش مصنوعی اگر در این راه استفاده شود یک تهدید برای کرامت انسانی محسوب می‌شود.

پاملامککوردوک در هنگام سخنرانی با زنان بیان کرد "من می‌خواهم شانسم را با یک کامپیوتر بی طرف امتحان کنم " که به این اشاره می‌کرد که قضاوت یک کامپیوتر به عنوان یک قاضی یا پلیس بی‌طرف است و غرایز انسانی در آن دیده نمی‌شود. بنیان‌گذار هوش مصنوعی جان مک کارتی نتیجه‌گیری اخلاقی ویزینبام را نقد می‌کند. "هنگامی که اخلاقیات انسان مبهم و ناکامل است، موجب می‌شود که گاهی استبداد را فراخواند".

بیل هیبارد می‌نویسد: «کرامت انسانی مستلزم آن است که ما برای خلاص شدن از جهل نسبت به موجودات جهان تلاش کنیم و هوش مصنوعی برای این تلاش لازم است»

شفافیت و متن باز

بیل هیبارد می‌گوید که به دلیل این که هوش مصنوعی تأثیر عمیقی بر انسانیت دارد، بنابراین توسعه دهندگان هوش مصنوعی نماینده انسانیت در آینده هستند و به همین دلیل باید تعهد اخلاقی آن‌ها در تلاش‌شان به وضوح مشخص باشد. بن گویرتزل و دیوید

هارت OpenCog را به عنوان یک منبع باز برای توسعه هوش مصنوعی ایجاد کردند. OpenAI یک شرکت تحقیقاتی هوش مصنوعی غیرانتفاعی ایجاد شده توسط ایلان ماسک، سام آلتمن و افرادی دیگر است که به منظور توسعه هوش مصنوعی به صورت منبع باز برای بهبود آینده بشریت فعالیت می‌کند. و همچنین تعداد بسیار زیادی شرکت‌های منبع باز دیگر برای این اهداف وجود دارند.

با تشکر از توجه شما

محمدصالح احتشامی نیا

۴۰۲۵۵۲۵۱۰۲۱