

# اخلاق ماشین: چالش‌ها و رویکردهای مسائل اخلاقی در هوش مصنوعی و ابرهوش

مجید رمضانی<sup>\*</sup>، دکتر محمدرضا فیضی درخشی<sup>۲</sup>

۱. گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی نبی اکرم (ص) تبریز

۲. گروه کامپیوتر، دانشکده کامپیوتر، دانشگاه تبریز

(تاریخ دریافت: ۹۱/۱۱/۵، تاریخ پذیرش: ۹۲/۴/۲۶)

## چکیده

**زمینه:** محققان تحقیق در عملیات در مورد نحوه انجام کار در سازمان‌ها به مدیران کمک می‌کنند و تصمیماتشان صرفاً بر مبنای کارایی و اثربخشی ارزیابی می‌شود. در محیط کسب‌وکار جهانی امروز، سازمان‌ها با مسائل پیچیده‌ای مواجه‌اند، به‌همین خاطر رویکردهای اخلاقی توجه محققان زیادی را به خود جلب کرده است. این مقاله فرایند نوآورانه و نسبتاً جدیدی با عنوان اخلاق گفتمانی را مورد ارزیابی قرار می‌دهد. رویکرد اخلاق گفتمانی با سیستم اخلاقیات سنتی تفاوت دارد. در سیستم اخلاقیات سنتی، تصمیمات اخلاقی بدون توجه به کمیته‌ها و مردمان واقعی درگیر در فرایند گفت‌وگو و تعامل مورد ارزیابی قرار می‌گیرد. در این مقاله با استفاده از نظریه اخلاق گفتمانی، جنبه‌های اخلاقی مطرح در مسائل گوناگون، معیار عملکرد ما برای مواجهه با این مسائل، و جهانی که این مسائل به آن تعلق دارند، تشریح شد. همچنین در ادامه برای مواجهه با هر یک از این نوع مسائل، روش‌شناسی‌های خاصی ارائه گردید.

**نتیجه‌گیری:** مسائل جهان امروز پیچیده و چندبعدی هستند. مدیران نمی‌توانند به جنبه‌های اخلاقی مسائل بی‌توجه باشند. اخلاق گفتمانی نظریه مناسبی برای تشریح جنبه‌های اخلاقی مسائل می‌باشد. تحقیق در عملیات برای مواجهه با چنین مسائلی بایستی آمادگی لازم را داشته باشد، و روش‌شناسی‌های مناسبی را ارائه نماید.

**کلید واژه‌ها:** اخلاق گفتمانی، تحقیق در عملیات سخت، تحقیق در عملیات نرم، روش‌شناسی انتقادی، روش‌شناسی ترکیبی

## سرآغاز

سبقت را از انسان ربوده‌اند. در سال ۱۹۹۸ محققین و صاحب نظران به تشریح این اندیشه پرداختند. آنها اظهار داشتند که رایانه‌ها ماشین‌های جامعی هستند که از توانایی انجام وظایف نامحدودی برخوردارند. هم چنین آنها روند پیشرفت این عرصه را از ابتدای پیدایش آن تا دورنمای آینده آن چنین تشبیه کردند؛ منظره‌ای از یک سرزمین وسیع شامل قابلیت‌های بشری تصور کنید، که دارای مناطق مسطحی با عناوین محاسبات ریاضی و حفظ طوطی‌وار مطالب، و تپه‌هایی با عناوین

از سال ۱۹۵۰ پیش‌تازان عرصه هوش مصنوعی<sup>۱</sup> همواره به رایانه (کامپیوتر)، به‌عنوان چشم ابزاری که به‌صورت بالقوه توانایی تفکر دارد نگریسته‌اند، و بر این عقیده بوده‌اند که روزی این ابزارها در وظایف هوشی از انسان پیشی خواهند گرفت، چنانچه در محاسبات منطقی بدون این که دچار خطا شوند، با در دسترس قرار دادن حجم انبوهی از فضای ذخیره‌سازی، گوی

\* نویسنده مسؤؤل: نشانی الکترونیکی: sir.ramezani@gmail.com

اثبات قضایا و شطرنج و نیز قتل مرتفعی با عناوین جابه‌جایی و حرکت، هماهنگی چشم و دست<sup>۲</sup> و تعاملات اجتماعی است. هر فردی در نقطه‌ای از این چشم انداز قرار دارد و برای دستیابی به نقاط دیگر نیازمند تلاش بسیاری است، و تنها برخی از افراد توانسته‌اند به چند نقطه دست پیدا کنند. عملکرد در حال پیشرفت رایانه، به آبی می‌ماند که به آرامی در این سرزمین در جریان است. حدود نیم قرن پیش این جریان مناطق مسطح را در بر گرفت، و ماشین حساب‌ها و کارمندان ثبت و ذخیره‌سازی اسناد را با خود برد. در عین حال بسیاری از انسان‌ها از این جریان مصون ماندند. امروزه جریان آب تپه‌ها را نیز در بر گرفته است، البته هنوز در قتل مرتفع انسان احساس امنیت دارد، اما به نظر می‌رسد با سرعت موجود این قله‌ها نیز در نیم قرن آینده در معرض این جریان قرار گرفته و زیر آب فرو روند. وقتی که بلندترین کوه‌ها با آب پوشیده شدند، ماشین‌هایی وجود خواهد داشت که همانند انسان در همه امور تعامل داشته باشند. ذهن موجود در این ماشین در آن هنگام خود آشکار<sup>۳</sup> خواهد بود (۱). امروزه دانش هوش مصنوعی تنها به دنبال ساخت ماشین‌هایی با عملکرد انسان گونه نیست، بلکه پیاده‌سازی ساختار درونی ذهن انسان را در دستور کار قرار داده است (۲ و ۳). اندیشه دست‌یابی به چنین ماشین‌هایی همواره عامل پیدایش سوالاتی در ذهن انسان بوده است، دسته اول سوالاتی هستند که در مورد امکان چنین رخدادی بحث می‌کنند. این سوالات به تعارض تاریخی میان طرفداران هوش مصنوعی ضعیف<sup>۴</sup> که معتقد به عملکرد هوشمند ماشین هستند و طرفداران هوش مصنوعی قوی<sup>۵</sup> که معتقد به امکان تفکر ماشین هستند، اشاره دارند (۴). دسته دوم نیز به مسایل اخلاقی این ماشین‌ها پرداخته‌اند. ایده اخلاق مصنوعی<sup>۶</sup> یا اخلاق ماشین<sup>۷</sup> (توانایی ماشین در تصمیم‌گیری اخلاقی) نیز یکی از جوانب عمده این پیشرفت محسوب می‌شود. البته با این که عامل‌های هوشمند<sup>۸</sup> پیشرفت قابل توجهی داشته‌اند، نظرات متعددی در این مورد وجود دارد. یکی از انعطاف‌ناپذیرترین دیدگاه‌ها بر این مبناست که رایانه فاقد هرگونه حالات اخلاقی است، چرا که قادر به درک معنای آنچه که مورد پردازش قرار می‌دهد نیست (۵). البته این بیان از دیدگاه فلسفی تایید می‌شود چرا که اطلاع و آگاهی یکی از

شرایط لازم در امر اخلاق<sup>۹</sup> محسوب می‌شود. در مقابل برخی نیز بر این باورند که انسان‌ها صلاحیت لازم برای زندگی با انتظارات اخلاقی را ندارند، و وظایف خطیر نظیر تصمیم‌گیری‌های اخلاقی یا قضاوت‌ها باید به عهده رایانه گذاشته شود (۶ و ۷). با این حال بیشتر تحقیقات با یک دیدگاه عملی بر این عقیده استوار هستند که عامل‌های مستقل<sup>۱۰</sup> (موجودیت‌های مستقلی که در راستای نیل به اهداف خود گام بر می‌دارند) دارای ویژگی‌های اخلاقی بوده و قادر به تصمیم‌گیری‌های اخلاقی هستند (۸-۱۱).

در این مقاله با هدف مطالعه لزوم بررسی مسائل اخلاقی مرتبط با ماشین و راهکارهای مواجهه با آن، ابتدا به تعریف اخلاق پرداخته و نیز اخلاق ماشین و دلیل توجه به آن مورد بررسی قرار خواهد گرفت. سپس به تبیین یکی از مهم‌ترین دلایل ناکامی هوش مصنوعی در راستای نیل به هدف غایی خود (که همان شبیه‌سازی رفتار جامع انسانی است) خواهیم پرداخت. در ادامه به معرفی ابرهوش<sup>۱۱</sup> و ویژگی‌های آن پرداخته و ضرورت بررسی مسائل اخلاقی در این ماشین‌ها مورد تاکید قرار خواهد گرفت. در انتها نیز با بررسی چالش‌های اخلاقی ابرهوش، راهکار مواجهه با مسائل اخلاقی این پدیده نوین مورد توجه قرار خواهد گرفت.

## اخلاق

در فرهنگ‌ها و کتب مختلف اخلاقی، تعاریف متفاوتی از اخلاق ارائه شده است، با این حال اختلاف عمیقی میان این تعاریف مشاهده نمی‌شود. اخلاق جمع واژه خُلق و به معنی خوی هاست (۱۲).

پیشینه مباحث اخلاقی به قدمت وجود آدمی است. اندیشیدن درباره معنای اخلاق با سقراط آغاز می‌شود. برخی از اندیشمندان اخلاق را عبارت از مجموعه قواعدی می‌دانند که به منظور رفع تعارضات درونی و بین اشخاص پدید آمده است (۱۳). به تعبیر ساده می‌توان اخلاق را شاخه‌ای از فلسفه دانست که در پی پاسخ‌گویی به سوالات قدیمی در مورد وظیفه،

درستی و صداقت، یکپارچگی، فضیلت، عدالت، زندگی نیک و نظایر آن است (۱۳).

شاخه نو ظهور اخلاق ماشین (یا ریاتیک) به دنبال تضمین رفتار اخلاق مدارانه ماشین است. به تعبیر دیگر این شاخه به دنبال این است که تضمین کند رفتار ماشین از لحاظ اخلاقی برای انسان قابل قبول است (۱۴). بطور کلی پاسخ به این سوال که آیا اخلاق ماشین وجود دارد، در گرو دست یابی به توافقاتی در مورد اخلاق ماشین است. برخی بر این باورند که حالت‌های اخلاقی برای ماشین به هیچ وجه وجود ندارد چرا که اخلاق یک امر احساسی است در حالی که ماشین فاقد احساس است. برخی نیز معتقدند که ماشین‌ها بدون تردید دارای حالت‌های اخلاقی هستند، با این استدلال که انسان خود یک ماشین بوده و دارای حالت‌های اخلاقی است (۱۵).

اخلاق ماشین از جوانب مختلفی قابل بحث و بررسی است و پژوهش در هر یک از این زمینه‌ها به سرعت به سمت مباحث عمیق و چالش برانگیز فلسفی سوق می‌یابد. علیرغم این پیچیدگی نباید در دنیای سرشار از فن آوری امروزی از توجه به مسائل اخلاقی ماشین سر باز زد. به علاوه این که به طور روز افزون وظایف تصمیم‌گیری بیشتری به ماشین محول می‌شود (مانند راننده خودکار خودرو).

### جنبه گم شده هوش مصنوعی

همه متخصصان دانش هوش مصنوعی بر این باورند که امروزه با این که الگوریتم<sup>۱۲</sup> های هوش مصنوعی توانسته اند در برخی از حوزه‌های خاص بر انسان پیروز شوند، اما بطور کلی قابل مقایسه با همه توانایی‌های انسان نیستند. برخی از متخصصین عقیده دارند که به محض این که متخصصین هوش مصنوعی دریافته‌اند که چگونه کاری را انجام دهند، در نظر گرفتن این توانایی به عنوان هوشمندی متوقف می‌شود؛ شطرنج همواره به عنوان نمادی از هوش شناخته می‌شد تا این که دیپ بلو<sup>۱۳</sup> (ماشین شطرنج بازی که توسط شرکت آی بی ام در سال ۱۹۹۷ ساخته شد) توانست کاسپاروف، قهرمان رقابت‌های جهانی شطرنج را شکست دهد. این متخصصین همچنین بر این

عقیده اند که برخی از جوانب هوش مصنوعی مفقود شده است (۱۶)، که عدم توجه به آن نقص عمده‌ای بر پیکره این دانش وارد کرده و مانع تحقق هدف آن خواهد بود. همه این افراد در مورد استفاده از عبارت هوش مصنوعی عمومی<sup>۱۴</sup> برای اشاره به هوش مصنوعی حقیقی اتفاق نظر دارند (۱۷). همان طور که پیداست عمومیت<sup>۱۵</sup> همان جنبه گم شده هوش مصنوعی است که عدم دست یابی به این جنبه عامل ناکامی این دانش در تحقق اهداف خود محسوب می‌شود. همه الگوریتم‌های هوش مصنوعی موجود که کارایی بهتر یا برابری نسبت به انسان دارند تنها در یک زمینه خاص کاربرد داشته و به تعبیر دیگر خاص منظوره هستند. دیپ بلو تنها قادر به انجام بازی شطرنج است و مهارت دیگری ندارد. چنین الگوریتم‌هایی نظیر بسیاری از گونه‌های حیات زیستی اطرافمان، به استثنای گونه انسان هستند. هر یک از حیوانات توانایی منحصر به فردی دارند، به عنوان مثال یک زنبور در ساخت کندو مهارت بسیاری دارد، و یک سگ آبی از توانایی خاصی در ساختن سد برخوردار است. اما هرگز زنبور توانایی ساخت سد و یا یک سگ آبی توانایی ساخت کندو را ندارد، در عین حال به نظر می‌رسد یک انسان توانایی هر دو کار را دارد. البته عمومیت هوش انسان خود قابل نزاع است، اما بطور یقین انسان در برخی از وظایف شناختی<sup>۱۶</sup> بسیار بهتر عمل کرده و هوش این گونه، بسیار پرکاربرد تر از دیگر گونه‌های حیات است (۱۸).

یکی دیگر از ویژگی‌هایی که در الگوریتم‌های هوش مصنوعی حتی در نوع خاص منظوره آنها بایستی وجود داشته باشد، توانایی برنامه ریزی غیر محلی<sup>۱۷</sup> آن هاست. در دامنه محلی رفتارهای سیستم بطور خاص تصور می‌شود. به عنوان مثال در الگوریتم دیپ بلو که قهرمان شطرنج جهان را شکست داد، در صورتی که برنامه نویسان همه فضای حالت بازی شطرنج را در یک پایگاه داده وارد کرده باشند، معیار غیر محلی بودن برنامه ریزی سیستم ارضا نشده و سیستم تنها قادر به انجام حرکت‌هایی خواهد بود که به آن دیکته شده است. البته این کاری نبود که توسط طراحان این الگوریتم انجام شد، چرا که اولاً فضای حالات بازی شطرنج بطور غیر قابل مدیریت بالا بوده و ثانیاً اگر همه ورودی‌هایی که در نظر طراحان خوب به نظر می‌رسد وارد

یاد می‌کند. مغزی که مانند انسان فکر می‌کند، اما خیلی سریع‌تر از آن (۲۲).

محققین مختلف اذعان دارند که احتمال عامل‌های ابرهوش که در دهه‌های آتی ایجاد خواهند شد، به اندازه کافی بالاست که این مساله را تحت توجه قرار دهد (۲۷-۲۳). البته همان‌طور که پیش‌تر ذکر شد وجود چنین عامل‌هایی و یا به تعبیر دیگر تحقق آرمان دانش هوش مصنوعی بطور چشم‌گیری در گرو دست‌یابی به جوانب گم‌شده این دانش است. حتی در صورت احتمال کم این رخداد، نتایجی که یک عامل ابرهوش در پی خواهد داشت به حدی زیاد است که توجه عمده‌ای را به خود می‌طلبد (۲۸ و ۲۹).

بررسی برخی از پیامدهای غیرعادی ابرهوش در درک بهتر موضوع موثر خواهد بود:

ابرهوش ممکن است آخرین ابداعی باشد که بشر نیاز دارد؛ از آنجا که این دستاورد توانایی‌های منطقی بالاتری نسبت به انسان دارد، عملکرد بسیار بهتری در تحقیقات علمی و پیشرفت فن‌آوری خواهد داشت، حتی ممکن است این برتری نسبت به کارهای گروهی انسان‌ها نیز مشهود باشد. یکی از پیامدهای صریح چنین رخدادی عبارت است از:

توسعه فن‌آوری در همه شاخه‌ها با پیدایش هوش مصنوعی پیشرفته، به سرعت رشد خواهد کرد؛ به احتمال زیاد فن‌آوری‌های قابل‌پیش‌بینی در زمینه‌های مختلف، به سرعت توسط اولین نسل از ابرهوش در دسترس قرار خواهد گرفت. بدون شک این اتفاق حتی برای بسیاری از فن‌آوری‌ها نیز که هیچ سرنخی از آنها در دست نیست، رخ خواهد داد.

ابرهوش خود منجر به تولید ابرهوش‌های پیشرفته‌تری خواهد شد؛ این جنبه هم از لحاظ سخت‌افزارها و هم از لحاظ نرم‌افزارهای بهبود داده شده توسط ابرهوش محقق خواهد شد. پیدایش ابرهوش ممکن است ناگهانی باشد؛ به نظر دست‌یابی به هوش مصنوعی انسان‌گونه، از جایی که اکنون درآنیم، بسیار سخت‌تر از دست‌یابی به ابرهوش از آن نقطه (هوش مصنوعی انسان‌گونه) است. به عبارت دیگر درحالی که دست‌یابی به هوش مصنوعی انسان‌گونه نیازمند مرور زمان است، مرحله‌نهایی و دست‌یابی به ابرهوش ممکن است به سرعت روی

سیستم شود، سیستم قادر به شکست دادن افراد ماهرتر از طراحان خود نخواهد بود. این در حالی است که طراحان، قهرمانان شطرنج نبوده و قادر به شکست کاسپاروف نیستند. در عوض طراحان این الگوریتم تصمیم گرفتند که حرکت‌های دیپ بلو طوری باشد که ملاک‌های غیر محلی بهینگی رعایت شود. یعنی همه حرکت‌ها در جهت هدایت آینده بازی و بردن آن است. چه بسا که در حرکت‌های محلی، هدف تنها یک حمله به شاه حریف خواهد بود (۱۹). لذا بطور کلی همه الگوریتم‌های هوش مصنوعی بایستی تضمینی از امنیت و موفقیت سیستم داشته باشند که این تنها در صورت عملکرد غیر محلی سیستم امکان‌پذیر است، بطوری که نتایج آتی سیستم قابل‌پیش‌بینی باشد.

## ابرهوش

محققین برای اولین بار در سال ۱۹۶۵ فرضیه ابرهوش را بنیان نهادند: یک سیستم هوشمند برای دست‌یابی به درک صحیحی از خود می‌تواند خود را دوباره طراحی کند، یا نمونه‌هایی از خود بسازد، که این نمونه‌ها می‌توانند بسیار هوشمندتر از سیستم‌های پیشین باشند. این سیستم‌ها نیز می‌توانند نمونه‌هایی از خود که هوشمندتر از خودشان هستند را تولید کنند، و همین‌طور تا آخر که یک چرخه بازخوردی مثبت<sup>۱۸</sup> ایجاد شود. این فرآیند انفجار هوش<sup>۱۹</sup> نام دارد (۲۰).

برای درک بهتر موضوع اجازه دهید به یک بررسی بپردازیم. افزایش سرعت پردازش گرهای بطور چشم‌گیری در ابرهوش تأثیرگذار است. سریع‌ترین اعصاب شناسایی شده تا کنون با سرعت ۱۰۰۰ بار در ثانیه آتش<sup>۲۰</sup> می‌کند. سریع‌ترین بندهای آکسون<sup>۲۱</sup> موجود با سرعت ۱۵۰ متر در ثانیه سیگنال‌ها را هدایت می‌کنند (۲۱). به نظر می‌رسد ساخت مغزی با سرعت یک میلیون برابر سرعت مغز انسان امکان‌پذیر است، بدون این که تغییری در اندازه آن بوجود آید. اگر ذهن انسان تسریع داده شود، در این صورت یک سال تفکر ذهنی<sup>۲۲</sup> می‌تواند تنها در ۳۱ ثانیه انجام شود و یک هزاره تنها هشت ساعت و نیم به طول انجامد. محققین از چنین مغز تسریع شده‌ای با عنوان ابرهوش ضعیف<sup>۲۳</sup>

در سال ۱۹۴۲ ایزاک آسیموف نویسنده روسی داستان‌های علمی-تخیلی به منظور جلوگیری از آسیب‌رسانی ربات‌ها به انسان، اقدام به تدوین سه قانون برای ربات‌های خود نمود که بعدها به قوانین رباتیک آسیموف مشهور شدند. بر اساس این قوانین (الف) یک ربات نباید با ارتکاب عملی یا خودداری از انجام عملی باعث آسیب دیدن یک انسان شود، (ب) یک ربات باید از فرامین انسان‌ها تبعیت کند مگر این که آن فرامین در تعارض با قانون نخست باشد، (ج) تا هنگامی که قانون نخست یا دوم زیر پا گذاشته نشده است ربات باید وجود خود را حفظ کرده و در بقای خود بکوشد. علاوه بر این که همه ربات‌ها در آثار آسیموف ملزم به رعایت این قوانین بودند، برخی از صنایع و نهادها نیز به تبعیت از آنها پرداختند (۳۲).

در حالی که سه قانون رباتیک آسیموف اساس کارآمدی برای تبیین اخلاق ماشین شناخته نشده‌اند، با این حال اندک توافقی میان محققین در مورد چگونگی ساختار اخلاقی عامل‌های مصنوعی اخلاقی وجود دارد (۳۱ و ۳۳). البته نظرها در این مورد در گستره وسیعی قرار دارند، شامل: استفاده از الگوریتم‌های تکاملی و ساخت جمعیتی از عامل‌های هوشمند برای دست‌یابی به یک نسل از «اخلاق‌مندترین عامل‌ها»، بر این اساس بدون این که عملکرد اخلاقی عامل‌ها از پیش تعیین شده باشد، امکان ترکیب مهارت‌های مختلف اخلاقی آنها در راستای دست‌یابی به اخلاق‌مندترین عامل‌ها فراهم می‌شود (۳۰). استفاده از مدل شناختی شبکه‌های عصبی<sup>۲۶</sup> که با استفاده از ابزار شبکه‌های عصبی مصنوعی و یک مجموعه آموزشی شامل حالت‌های مختلف اخلاقی ماشین برای آموزش سیستم، امکان نتیجه‌گیری در حالت‌های مختلف اخلاقی برای عامل‌ها فراهم می‌شود؛ و استفاده از روش‌های مختلف ترکیبی که شامل ترکیب روش‌های مختلف دست‌یابی به عامل‌های اخلاقی است؛ تا استفاده از نظام ارزش‌ها<sup>۲۷</sup> که مبنی بر قواعد طلایی<sup>۲۸</sup> بوده و شامل مجموعه‌ای از ارزش‌های اخلاقی برای ماشین است و هم‌چنین استفاده از فضیلت اخلاق<sup>۲۹</sup> که با تاکید بر رفتار عامل‌ها و نیز ارزش‌های اخلاقی موجود در آنها به دنبال تصمیم‌گیری در مورد مسائل اخلاقی مختلف است (۳۰ و ۳۴ و ۳۵).

دهد. این رویداد انتقالی خواهد بود از یک هوش مصنوعی تقریباً انسان‌گونه به یک ابرهوش تمام‌عیار. با استفاده از فرآیندهای تکاملی سرعت این فرآیند افزایش خواهد یافت، شاید یک روز بجای یک سال. احتمال پیدایش ناگهانی ابرهوش به فرضیه تفرّد<sup>۳۴</sup> ارجاع دارد (بر اساس این فرضیه بشر روزی به ماشین‌هایی دست پیدا خواهد کرد که با استفاده از هوش مصنوعی بسیار هوشمندتر از انسان‌ها عمل خواهند کرد) (۲۲).

عقل‌های مصنوعی بطور بالقوه عامل‌های مستقلی هستند؛ در عین حال که اختصاص ابرهوش برای مسائل خاص ممکن است اما نباید از آن به عنوان یک ابزار محض تصور شود. یک ابرهوش جامع مستقلاً قادر به ابتکار بوده و توانایی برنامه‌ریزی طرح‌های خود را دارد. لذا چنین عامل‌هایی می‌توانند کاملاً مستقل در نظر گرفته شوند.

عقل‌های مصنوعی مقید به محرک‌های انسان‌گونه نیستند؛ انسان‌ها به ندرت تمایل به بردگی و خدمت به دیگری را دارند، اما این که هدف یک ابرهوش خدمت به بشر یا یک انسان خاص باشد، کاملاً شدنی است. بدون این که هیچ تلاشی در پی شورش و آزادی خود داشته باشد. این یعنی این که امکان ساخت ابرهوش‌هایی با اهداف خاص برای انجام فعالیت‌هایی که برای انسان نیازمند انگیزه‌های مختلف هستند وجود دارد. خوب یا بد، عقل‌های مصنوعی نیازمند تاثیر از گرایش‌های انگیزه‌ای انسان‌ها نیستند.

## چالش‌های اخلاقی ابرهوش

حوزه تحقیقاتی اخلاق ماشین اخیراً به عنوان یک زیرشاخه از هوش مصنوعی ظهور پیدا کرد که در مورد تضمین رفتارهای اخلاقی عامل‌های هوشمند به بحث و تبادل نظر پرداخته و دانشمندان حوزه فلسفه و علوم کامپیوتر را به مشارکت دعوت کرده است. از این حوزه هم‌چنین با نام عامل‌های مصنوعی اخلاقی<sup>۲۵</sup> نیز یاد می‌شود (۳۰). با تمرکز روی رفتار عامل‌های مصنوعی، این شاخه از شاخه اخلاق تکنولوژی که در مورد نحوه استفاده تکنولوژی توسط بشر بحث می‌کند، متمایز می‌گردد (۳۱).

فرض: عامل‌های مصنوعی اخلاقی توسط انسان ساخته شده‌اند و نمی‌توانند معماری خود را تغییر دهند (۳۷).  
چالش طراحی: در این صورت اهداف سیستم‌های هوش مصنوعی بایستی ثابت باشد، به عبارت دیگر عامل‌ها نباید درصدد تغییر اهداف خود به اهداف دیگری باشند (۲۵ و ۳۷)، که این امر اساساً در تعارض با تعریف ارائه شده برای ابرهوش است. فرض: انسان بسیار قدرتمندتر از عامل‌های مصنوعی اخلاقی است لذا استفاده از تئوری‌های بازی برای اجبار عامل به همکاری امکان‌پذیر است.

چالش طراحی: این فرض نیز با نادیده گرفتن توانایی‌های ابرهوش در تعارض با تعریف ارائه شده برای آن است. بعلاوه این که عامل‌های مصنوعی اخلاقی بایستی مانند نوع بشر پیامدهای خودخواهانه را ترجیح داده و ارجحیت خود را تحمیل کنند. توانایی منطقی بالای این عامل‌ها در مقایسه با انسان، مانع از تحقق این فرض خواهد شد.

فرض: عامل‌های مصنوعی اخلاقی می‌توانند در محیط‌هایی که بسیار به محیط طراحی خود شبیه هستند مورد تست و آزمایش قرار گیرند (۳۰).

چالش طراحی: این فرض نیز با تحدید کاربردهای عامل‌های هوشمند و نادیده گرفتن توانایی‌های آن‌ها، در راستای ماشین‌های موقعیت-کنش گام برداشته و در تعارض با مفهوم واقعی ابرهوش است. با افزایش پیچیدگی‌های هوش مصنوعی و پیشرفت آن، دانش آن‌ها، محیط، توانایی‌ها و محرک‌ها تغییر خواهند کرد، لذا سیستم‌هایی که پیش‌تر تولید شده‌اند ممکن است دارای رفتارهایی باشند که برای بشر امروزی پر مخاطره باشد (۲۵). به تعبیر دیگر در صورتی که عامل‌های مصنوعی اخلاقی تنها برای محیط‌هایی با محرک‌ها و شرایط خاص طراحی شده باشند، تضمینی از رفتار امن آنها با قرارگیری در محیط‌هایی با محرک‌ها و شرایط متفاوت، وجود نخواهد داشت. فرض: طراحی عامل‌های مصنوعی اخلاقی یک فرآیند افزایشی و تکراری است، که فرصت نظارت انسانی را فراهم می‌کند (۳۰).

چالش طراحی: با توجه به توانایی‌های منطقی بالای عامل‌های هوشمند همان‌طور که پیش‌تر ذکر شد، دست‌یابی به ابرهوش

چگونه یک عامل اخلاقی مصنوعی می‌تواند ارزش‌های انسانی را نمایش دهد؟ در سال ۲۰۰۶ محققین و صاحب‌نظران در راستای دست‌یابی به یک رویکرد عملی برای مواجهه با مسائل اخلاقی ماشین به تشریح دو کلاس از عامل‌ها پرداختند: ماشین‌های موقعیت-کنش<sup>۳۰</sup> که دارای قوانینی برای تبیین فعالیت لازم در پاسخ به یک محرک خاص هستند (در این ماشین‌ها همه حالت‌ها و همچنین کنش مناسب ماشین در هر یک از این حالت‌ها پیش‌بینی شده است) و ماشین‌های انتخابی<sup>۳۱</sup> که قادر به بهینه‌سازی پیامدهای ممکن بوده و می‌توانند فعالیت مربوط به بهینه‌ترین خروجی را انتخاب کنند. دسته اول از ماشین‌ها می‌توانند رفتارهای پیچیده‌ای از خود نشان دهند، اما از آنجا که تنها اهداف ضمنی<sup>۳۲</sup> دارند، رفتارهای غیر قابل انعطافی داشته و به سختی در محیط‌ها و شرایط جدید قابل استفاده هستند. در مقایسه، دسته دوم به راحتی در هر شرایط غیر قابل انتظاری، مبتنی بر ارزش‌ها و اهداف روشنی که در آنها تبیین شده است، بهترین عملکرد را انتخاب می‌کنند (۳۶).

محققان اخلاق ماشین به این که هر عامل مصنوعی اخلاقی می‌تواند یک عامل اخلاقی ضمنی باشد اتفاق نظر دارند، که می‌تواند همه اهداف مورد انتظار خود را تحت یک رفتار کاملاً امن برآورده سازد، بدون اینکه نیازی به بسط استدلال اخلاقی به شرایط جدید باشد (۳۳). به عبارت دیگر بیشتر مباحث موجود در اخلاق ماشین تلویحاً بر این فرض استوارند که ویژگی‌های کلیدی معینی در عامل‌های اخلاقی، ثابت در نظر گرفته شوند و توسعه در این زمینه به سمت تشکیل اهداف ضمنی برای ماشین‌ها و دست‌یابی به ماشین‌های موقعیت-کنش گام بر دارد. بایستی اذعان داشت که اتخاذ این فرایض، شباهت کمی به کاربرد عامل‌های ابرهوش داشته و حتی معیار غیر محلی بودن برنامه ریزی سیستم نیز ارضا نمی‌شود. هر یک از این فرایض به نوعی با تعاریف و بایدهای مفهوم ابرهوش در تعارض بوده و چالش‌هایی را در پی دارند. البته بایستی توجه داشت که با این حال حذف هر یک از آنها نیز چالش‌های طراحی جدیدی را می‌طلبد. برخی از این فرایض و چالش‌های پیش‌روی آنها عبارتند از:

تحقق آرمان‌های هوش مصنوعی، شاهد نسل جدیدی از ماشین‌ها تحت عنوان ابرهوش خواهیم بود که قادر به طراحی ماشین‌هایی هوشمندتر از خود هستند. این ماشین‌ها نیز می‌توانند نمونه‌هایی هوشمندتر از خود ساخته و همین‌طور تا آخر در یک چرخه بازخوردی مثبت شاهد طراحی ماشین‌هایی هوشمندتر از مرحله قبل باشیم. از این واقعه تحت عنوان انفجار هوش یاد می‌شود.

با افزایش روز افزون وظایف تصمیم‌گیری ماشین و نیز برتری عقلانی نسل جدید ماشین‌های تولید شده، به منظور تضمین رفتار امن ماشین‌ها توجه با مسائل اخلاقی این ماشین‌ها امری ضروری خواهد بود. در این راستا دو رویکرد برای طراحی ماشین‌های هوشمند پیشنهاد شده است که عبارتند از ماشین‌های موقعیت-کنش و ماشین‌های انتخابی. در نوع اول از ماشین‌ها با پیش‌بینی همه حالت‌های ممکن به تبیین کنش‌های مناسب ماشین پرداخته شده است در حالی که در نوع دوم با در دسترس بودن مجموعه‌ای از کنش‌ها، ماشین قادر به انتخاب مناسب‌ترین کنش و بهینه‌سازی خروجی است. علیرغم این که ماشین‌های موقعیت-کنش در تعارض با مفهوم ابرهوش هستند، اما برای تضمین رفتار امن ماشین‌ها با بایستی در این راستا گام برداشته شود.

ممکن است ناگهانی باشد. همچنین با توجه به این که ابرهوش ممکن است آخرین ابداعی باشد که بشر نیاز دارد، این عامل قادر خواهد بود با سرعت بالا و توانایی‌های چشم‌گیر خود، عامل‌هایی به مراتب توانمندتر از خود ایجاد کند. به تعبیر دیگر پتانسیل موجود در انفجار هوش می‌تواند منجر به پیشرفت‌های آشکاری در میزان افزایش هوش شود و عامل ابرهوش را برای وضع هر هدفی آزاد بگذارد (۲۵). لذا با توجه به تعریف ارائه شده برای ابرهوش و هم‌چنین توانایی‌های آن، امکان نظارت انسانی در طراحی این عامل‌ها میسر نخواهد بود.

همان‌طور که مشاهده می‌شود هر یک از این فرایض که به نحوی در جهت اهداف ماشین‌های موقعیت-کنش اتخاذ شده‌اند، به نوعی در تعارض با مفهوم ابرهوش هستند.

برای مقابله با چالش‌های مذکور و تحقق مفهوم ابرهوش، سیستم‌ها بایستی به سمت ماشین‌های انتخابی گام برداشته و یک معماری که در بردارنده ارزش‌های انسانی باشد، طراحی شود. از آنجا که ما هنوز برخی از دلایل درونی ادراکات و اندیشه‌های اخلاقی انسان را نمی‌دانیم، طراحی چنین سیستمی بدون توجه به محدودیت‌های ابزاری، بسیار پیچیده خواهد بود (۱۹ و ۳۸).

## نتیجه‌گیری

با پیشرفت دانش هوش مصنوعی و موفقیت آن در شبیه‌سازی برخی رفتارهای الگوی انسانی، حتی با حذف برخی از محدودیت‌های لاجرم موجود در آن نظیر خستگی و امکان اشتباه، شاهد ظهور ماشین‌هایی هستیم که حتی بهتر از الگوی انسانی قادر به انجام وظایف خود هستند. علیرغم این موفقیت‌ها بایستی اذعان داشت که هنوز دست‌یابی به هدف غایی هوش مصنوعی (شبیه‌سازی رفتار جامع انسانی) میسر نشده است. یکی از مهم‌ترین موانع دست‌یابی به این هدف عدم وجود عمومیت در وظایف عامل‌های هوشمند است. به عبارت دیگر موفقیت‌های کسب شده برای دانش هوش مصنوعی تنها در زمینه‌های خاص و تک‌منظوره بوده و به‌طور کلی می‌توان شاهد برتری انسان نسبت به ماشین بود. با رفع موانع موجود و

## واژه‌نامه

1. Artificial Intelligence (AI)	هوش مصنوعی
2. Hand-eye Coordination	هماهنگی چشم و دست
3. Self-Evident	خود آشکار
4. Weak AI	هوش مصنوعی ضعیف
5. Strong AI	هوش مصنوعی قوی
6. Artificial Ethics	اخلاق مصنوعی
7. Machine Ethics	اخلاق ماشین
8. Intelligent Agent	عامل هوشمند
9. Ethics	اخلاق
10. Autonomous Agent	عامل مستقل
11. Superintelligence	ابرهوش
12. Algorithm	الگوریتم
13. Deep Blue	دیپ بلو

- Experimental & Theoretical Artificial Intelligence 12(3):251-261.
10. Gotterbarn D (2002). The ethical computer grows up: Automating ethical decisions. In Alvarez et al. (Eds.). In Proceedings of the sixth ETHICOMP conference. Lisbon, Portugal. pp: 125-141.
11. Mowbray M (2002). Ethics for bots. 14<sup>th</sup> International Conference on System Research, Informatics and Cybernetics. vol. 1. Baden-Baden. pp: 24-28.
۱۲. آهنگری، فرشته «پیشینه و بنیادهای اخلاق در ایران و جهان» فصلنامه اخلاق در علوم و فناوری، ۱۳۸۶، شماره ۳ و ۴: ۱۱-۲۲
۱۳. ایمان، محمدتقی، غفاری نسب، اسفندیار «معیارهای اخلاقی در پژوهش‌ها ی علوم انسانی» فصلنامه اخلاق در علوم و فناوری، ۱۳۹۰، شماره ۲: ۶۶-۷۵.
14. Anderson SA (2011). Machine Metaethics. 1<sup>st</sup> Edition. Cambridge University Press. Cambridge. pp: 21-27.
15. Moor JH (2011). The Nature, Importance, and Difficulty of Machine Ethics. Machine Metaethics. 1<sup>st</sup> Edition. Cambridge University Press. Cambridge. pp: 13-20.
16. Hofstadter D (2006). Trying to Muse Rationally about the Singularity Scenario. Presented at the Singularity Summit at Stanford.
17. Schmidhuber J. Thorisson KR, Looks M (2011) Artificial General Intelligence First Edition springer verlag New York, pp. 1-10.
18. Hirschfeld LA, Gelman SA (1994). Mapping the Mind: Domain Specificity in Cognition and Culture. 1<sup>st</sup> Edition., Cambridge University Press. Cambridge. pp: 85-110.
19. Haidt J (2000). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review. 104:814-834.
20. Good IJ (1965). Speculations Concerning the First Ultraintelligent Machine. First Edition Academic Press New York pp. 31-88.
21. Sandberg A (1999). The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains. Journal of Evolution and Technology. 5:7-18.
22. Vinge V (1993). The Coming Technological singularity singularity. Presented at the VISION-21 symposium.
23. Bostrom N (2003). Ethical Issues In Advanced Artificial Intelligence. 2<sup>nd</sup> Edition. International Institute of Advanced Studies in Systems Research and Cybernetics. Baden-Baden. pp: 12-17.
24. Moravec H (1999). Robot: Mere Machine to Transcendent Mind. 1<sup>st</sup> Edition. Oxford University Press. New York. pp: 15-26.
25. Yudkowsky E (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk.
14. Artificial General Intelligence (AGI) هوش مصنوعی عمومی
15. Generality عمومیت
16. Cognitive شناختی
17. Nonlocal غیر محلی
18. Feedback Cycle چرخه بازخوردی
19. Intelligence Explosion انفجار هوش
20. Fire آتش کردن
21. Axon آکسون
22. Subjective ذهنی
23. Weak Superintelligence ابرهوش ضعیف
24. Singularity Hypothesis فرضیه تفرّد
25. Artificial Moral Agent عامل اخلاقی مصنوعی
26. Neural Networks شبکه های عصبی
27. Value Systems نظام ارزشی
28. Golden Rules قواعد طلایی
29. Virtue Ethics فضیلت اخلاق
30. Situation-Action Machine ماشین موقعیت-کنش
31. Choice Machine ماشین انتخابی
32. Implicit ضمنی

## منابع

- Moravec H (1998). When will computer hardware match the human brain. Journal of Evolution and Technology. 1: 1-12.
- The Blue Brain Project website. Available at: <http://bluebrain.epfl.ch.htm>. Accessed: 25 Nov. 2012.
- The Institute of the Ecole Polytechnique in Lausanne website. Available at: <http://www.epfl.ch/index.fr.html>. Accessed: 25 Nov. 2012.
- Russel S, Norvig P (2010). Artificial Intelligence: A modern approach. 3<sup>rd</sup> edition. Pearson Education. New Jersey. pp: 26-58.
- Stahl BC (2004). Information, Ethics, and Computers: The problem of autonomous moral agents. Minds and Machines 14: 67-83.
- Bechtel W (1985). Attributing responsibility to computer systems. Metaphilosophy, 16: 296-306.
- Stewart I (1996). The interrogator's fallacy. Scientific American. 275: 172-175.
- Allen C (2002). Calculated morality: Ethical computing in the limit. In Smit & Lasker (Eds.). Cognitive, emotive and ethical aspects of decision making and human action. vol I. Germany/Windsor. Ontario: Baden Baden. pp: 19-23.
- Allen C, Varner G, Zinser J (2000). Prolegomena to any future artificial moral agent Journal of

- In Global Catastrophic Risks, Bostrom and Cirkovic, Eds. Oxford University Press. Oxford. pp: 308-345.
26. Kurzweil R (2005). The Singularity Is Near. 1<sup>st</sup> Edition. Penguin Books. London. pp: 1-12.
  27. Hall JS (2007). Beyond AI: Creating the Conscience of the Machine. 1<sup>st</sup> Edition. Prometheus Books. US. pp: 45-59.
  28. Posner R (2004). Catastrophe: Risk and Response. 1<sup>st</sup> Edition. Oxford University Press. Oxford. pp: 16-83.
  29. Rees M (2004). Our Final Hour: A Scientist's Warning: how Terror, Error and Environmental Disaster Threaten Humankind's Future in this Centure - on Earth and Beyond. 1<sup>st</sup> Edition. Basic Books. New York. pp: 15-25.
  30. Wallach W, Collin A (2005). Android Ethics: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties. Proceedings of the 2005 COGSCI workshop: Toward Social Mechanics of Android Science. pp. 149-159.
  31. Anderson M, Anderson SL (2007). The status of machine ethics: a report from the AAAI Symposium. Minds and Machines. 17: 1-10.
  32. Joseph A, Angelo Jr (2006). Robotics: A Reference Guide to the New Technology. First Edition. Greenwood Press. Westport (US). pp: 18-26.
  33. Weng YH, Chen CH, Chuen T (2009). Toward the Human-Robot Co-Existence Society: On Safety Intelligence for Next Generation Robots. International Journal of Social Robotics. 1: 267-282.
  34. Guarini M (2005). Particularism and Generalism: How AI can Help us to Better Understand. Technical Report for Machine Ethics Symposium. American Association for Artificial Intelligence Fall Symposium, November. University of Windsor. Canada.
  35. Anderson M, Anderson SL (2007). The Consequences for Human Beings of Creating Ethical Robots. Proceedings of AAAI Workshop Human Implications of Human-Robot Interaction. Vancouver. British Columbia. Canada.
  36. Drescher G (2006). Good and Real: Demystifying Paradoxes from Physics to Ethics. 1<sup>st</sup> Edition. MIT Press. Massachusetts. pp: 35-221.
  37. Omohundro SM (2007). The Nature of Self-Improving Artificial Intelligence. Presented and distributed at Singularity Summit at San Francisco.
  38. Greene J (2002). The Terrible, Horrible, No Good, Very Bad Truth about Morality and What to Do About it. Thesis. Princeton University, US, Nov. 2002.