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

مجید رمضانی*[1]، دکتر محمدرضا فیضی درخشی[2]

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

از سال ۱۹۵۰ پیشتازان عرصه هوش مصنوعی[1] 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 ذخیره سازی، گوی سبقت را از انسان ربوده‌اند. در سال ۱۹۹۸ محققین و صاحب نظران به تشریح این اندیشه پرداختند. آنها اظهار داشتند که رایانه‌ها ماشین های 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؛ منظره ای 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

* نویسنده مسؤول: نشانی الکترونیکی: sir.ramezani@gmail.com

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

اثبات قضایا و شطرنج و نیز قلل مرتفعی با عناوین جابه‌جایی و حرکت، هماهنگی چشم و دست[۲] و تعاملات اجتماعی است. هر فردی در نقطه ای از این چشم انداز قرار دارد و برای دست‌یابی به نقاط دیگر نیازمند تلاش بسیاری است، و تنها برخی از افراد توانسته اند به چند نقطه دست پیدا کنند. عملکرد در حال پیشرفت رایانه، به آبی می ماند که به آرامی در این سرزمین در جریان است. حدود نیم قرن پیش این جریان مناطق مسطح را در بر گرفت، و ماشین حساب ها و کارمندان ثبت و ذخیره سازی اسناد را با خود برد. در عین حال بسیاری از انسان ها از این جریان مصون ماندند. امروزه جریان آب تپه ها را نیز در بر گرفته است، البته هنوز در قلل مرتفع انسان احساس امنیت دارد، اما به نظر می رسد 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، ماشین هایی وجود خواهد داشت که همانند انسان در همه امور تعامل داشته باشند. ذهن موجود در این ماشین در آن هنگام خود آشکار[۳] خواهد بود (۱).

امروزه دانش هوش مصنوعی تنها به دنبال ساخت ماشین هایی با عملکرد انسان گونه نیست، بلکه پیاده سازی ساختار درونی ذهن انسان را در دستور کار قرار داده است (۲و ۳). اندیشه دست یابی به چنین ماشین هایی همواره عامل پیدایش سوالاتی در ذهن انسان بوده است، دسته اول سوالاتی هستند که در مورد امکان چنین رخدادی بحث می کنند. این سوالات به تعارض تاریخی میان طرفداران هوش مصنوعی ضعیف[۴] که معتقد به عملکرد هوشمند ماشین هستند و طرفداران هوش مصنوعی قوی[۵] که معتقد به امکان تفکر ماشین هستند، اشاره دارند (۴). دسته دوم نیز به مسایل اخلاقی این ماشین ها پرداخته اند. ایده اخلاق مصنوعی[۶] یا اخلاق ماشین[۷] (توانایی ماشین در تصمیم گیری اخلاقی) نیز یکی از جوانب عمده این پیشرفت محسوب می شود. البته با این که عامل‌های هوشمند[۸] پیشرفت قابل توجهی داشته اند، نظرات متعددی در این مورد وجود دارد. یکی از انعطاف ناپذیرترین دیدگاه‌ها بر این مبناست که رایانه فاقد هرگونه حالات اخلاقی است، چرا که قادر به درک معنای آنچه که مورد پردازش قرار می دهد نیست (۵). البته این بیان از دیدگاه فلسفی تایید می شود چرا که اطلاع و آگاهی یکی از شرایط لازم در امر اخلاق[۹] محسوب می شود. در مقابل برخی نیز بر این باورند که انسان‌ها صلاحیت لازم برای زندگی با انتظارات اخلاقی را ندارند، و وظایف خطیر نظیر تصمیم گیری‌های اخلاقی یا قضاوت‌ها باید به عهده رایانه گذاشته شود (۶و ۷). با این حال بیشتر تحقیقات با یک دیدگاه عملی بر این عقیده استوار هستند که عامل‌های مستقل[۱۰] (موجودیت‌های مستقلی که در راستای نیل به اهداف خود گام بر می‌دارند) دارای ویژگی‌های اخلاقی بوده و قادر به تصمیم‌گیری‌های اخلاقی هستند (۱۱-۸).

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 مهم ترین دلایل ناکامی هوش مصنوعی در راستای نیل به هدف غایی خود (که همان شبیه سازی رفتار جامع انسانی است) خواهیم پرداخت. در ادامه به معرفی ابرهوش[۱۱] و ویژگی های آن پرداخته و ضرورت بررسی مسائل اخلاقی در این ماشین‌ها مورد تاکید قرار خواهد گرفت. در انتها نیز با بررسی چالش های 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 نمی شود. 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 می شود. برخی از اندیشمندان اخلاق را عبارت از مجموعه قواعدی می دانند که به منظور رفع تعارضات درونی و بین اشخاص پدید آمده است (۱۳). به تعبیر ساده می توان اخلاق را شاخه ای از فلسفه دانست که در پی پاسخ گویی به سوالات قدیمی در مورد وظیفه،

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

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

شاخه نو ظهور اخلاق ماشین (یا رباتیک) 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 زمینه ها به سرعت به سمت مباحث عمیق و چالش برانگیز فلسفی سوق می یابد. 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 تصمیم گیری بیشتری به ماشین محول می شود (مانند راننده خودکار خودرو).

جنبه گم شده هوش مصنوعی

همه متخصصان دانش هوش مصنوعی بر این باورند که امروزه با این که الگوریتم[۱۲] 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 دریافتند که چگونه کاری را انجام دهند، در نظر گرفتن این توانایی به عنوان هوشمندی متوقف می شود؛ شطرنج همواره به عنوان نمادی از هوش شناخته می شد تا این که دیپ بلو[۱۳] (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 عمده ای بر پیکره این دانش وارد کرده و مانع تحقق هدف آن خواهد بود. همه این افراد در مورد استفاده از عبارت هوش مصنوعی عمومی[۱۴] برای اشاره به هوش مصنوعی حقیقی اتفاق نظر دارند (۱۷). همان طور که پیداست عمومیت[۱۵] همان جنبه گم شده هوش مصنوعی است که عدم دست یابی به این جنبه عامل ناکامی این دانش در تحقق اهداف خود محسوب می شود. همه الگوریتم های 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 الگوریتم هایی نظیر بسیاری از گونه های 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 می رسد 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[۱۶] بسیار بهتر عمل کرده و هوش این گونه، بسیار پرکاربرد تر از دیگر گونه‌های حیات است (۱۸).

یکی دیگر از ویژگی هایی که در الگوریتم های هوش مصنوعی حتی در نوع خاص منظوره آنها بایستی وجود داشته باشد، توانایی برنامه ریزی غیر محلی[۱۷] آن هاست. در دامنه محلی رفتارهای سیستم بطور خاص تصور می شود. 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 حرکت هایی خواهد بود که به آن دیکته شده است. البته این کاری نبود که توسط طراحان این الگوریتم انجام شد، چرا که اولا فضای حالات بازی شطرنج بطور غیر قابل مدیریتی بالا بوده و ثانیا اگر همه ورودی هایی که در نظر طراحان خوب به نظر می رسد وارد

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

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 حرکت های 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 الگوریتم های هوش مصنوعی بایستی تضمینی از امنیت و موفقیت سیستم داشته باشند که این تنها در صورت عملکرد غیر محلی سیستم امکان پذیر است، بطوری که نتایج آتی سیستم قابل پیش بینی باشد.

ابرهوش

محققین برای اولین بار در سال ۱۹۶۵ فرضیه ابرهوش را بنیان نهادند: یک سیستم هوشمند برای دست یابی به درک صحیحی از خود می تواند خود را دوباره طراحی کند، یا نمونه هایی از خود بسازد، که این نمونه‌ها می توانند بسیار هوشمندتر از سیستم‌های پیشین باشند. این سیستم‌ها نیز می توانند نمونه هایی از خود که هوشمندتر از خودشان هستند را تولید کنند، و همین طور تا آخر که یک چرخه بازخوردی مثبت[18] ایجاد شود. این فرآیند انفجار هوش[19] نام دارد (۲۰).

برای درک بهتر موضوع اجازه دهید به یک بررسی بپردازیم. افزایش سرعت پردازش گرها بطور چشم گیری در ابرهوش تاثیرگذار است. سریع ترین اعصاب شناسایی شده تا کنون با سرعت ۱۰۰۰ بار در ثانیه آتش[20] می کند. سریع ترین بندهای آکسون[21] موجود با سرعت ۱۵۰ متر در ثانیه سیگنال‌ها را هدایت می کنند (۲۱). به نظر می رسد ساخت مغزی با سرعت یک ملیون برابر سرعت مغز انسان امکان پذیر است، بدون این که تغییری در اندازه آن بوجود آید. اگر ذهن انسان تسریع داده شود، در این صورت یک سال تفکر ذهنی[22] می تواند تنها در ۳۱ ثانیه انجام شود و یک هزاره تنها هشت ساعت و نیم به طول انجامد. محققین از چنین مغز تسریع شده ای با عنوان ابرهوش ضعیف[23] یاد می کنند. مغزی که مانند انسان فکر می کند، اما خیلی سریع تر از آن (۲۲).

محققین مختلف اذعان دارند که احتمال عامل‌های ابرهوش که در دهه های آتی ایجاد خواهند شد، به اندازه کافی بالاست که این مساله را تحت توجه قرار دهد (۲۷- ۲۳). البته همان طور که پیش تر ذکر شد وجود چنین عامل هایی و یا به تعبیر دیگر تحقق آرمان دانش هوش مصنوعی بطور چشم گیری در گرو دست یابی به جوانب گم شده 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 عمده ای را به خود می طلبد (۲۸و۲۹).

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 فن آوری 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 فن آوری در همه شاخه‌ها با پیدایش هوش مصنوعی پیشرفته، به سرعت رشد خواهد کرد؛ به احتمال زیاد فن آوری‌های قابل پیش بینی در زمینه‌های مختلف، به سرعت توسط اولین نسل از ابرهوش در دسترس قرار خواهد گرفت. بدون شک این اتفاق حتی برای بسیاری از فن آوری‌ها نیز که هیچ سرنخی از آنها در دست نیست، رخ خواهد داد.

ابرهوش خود منجر به تولید ابرهوش‌های پیشرفته تری خواهد شد؛ این جنبه هم از لحاظ سخت افزارها و هم از لحاظ نرم افزارهای بهبود داده شده توسط ابرهوش محقق خواهد شد. پیدایش ابرهوش ممکن است ناگهانی باشد؛ به نظر دست یابی به هوش مصنوعی انسان گونه، از جایی که اکنون درآنیم، بسیار سخت تر از دست یابی به ابرهوش از آن نقطه (هوش مصنوعی انسان گونه) است. به عبارت دیگر درحالی که دست یابی به هوش مصنوعی انسان گونه نیازمند مرور زمان است، مرحله نهایی و دست یابی به ابرهوش ممکن است به سرعت روی

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

دهد. این رویداد انتقالی خواهد بود از یک هوش مصنوعی تقریبا انسان گونه به یک ابرهوش تمام عیار. با استفاده از فرآیندهای تکاملی سرعت این فرآیند افزایش خواهد یافت، شاید یک روز بجای یک سال. احتمال پیدایش ناگهانی ابرهوش به فرضیه تفرّد[۲۴] ارجاع دارد (بر اساس این فرضیه بشر روزی به ماشین هایی 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 برنامه ریزی طرح‌های خود را دارد. لذا چنین عامل هایی می توانند کاملا مستقل در نظر گرفته شوند.

عقل‌های مصنوعی مقید به محرک‌های انسان گونه 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 ابرهوش هایی با اهداف خاص برای انجام فعالیت هایی که برای انسان نیازمند انگیزه های مختلف هستند وجود دارد. خوب یا بد، عقل‌های مصنوعی نیازمند تاثیر از گرایش‌های انگیزه ای انسان‌ها نیستند.

چالش‌های اخلاقی ابرهوش

حوزه تحقیقاتی اخلاق ماشین اخیرا 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[۲۵] نیز یاد می شود(۳۰). با تمرکز روی رفتار عامل‌های مصنوعی، این شاخه از شاخه اخلاق تکنولوژی که در مورد نحوه استفاده تکنولوژی توسط بشر بحث می کند، متمایز می گردد(۳۱).

در سال ۱۹۴۲ ایزاک آسیموف نویسنده روسی داستان‌های علمی–تخیلی به منظور جلوگیری از آسیب رسانی 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 نشده اند، با این حال اندک توافقی میان محققین در مورد چگونگی ساختار اخلاقی عامل‌های مصنوعی اخلاقی وجود دارد(۳۱و ۳۳). البته نظرها در این مورد در گستره وسیعی قرار دارند، شامل: استفاده از الگوریتم‌های تکاملی و ساخت جمعیتی از عامل‌های هوشمند برای دست یابی به یک نسل از «اخلاقمندترین عامل ها»، بر این اساس بدون این که عملکرد اخلاقی عامل‌ها از پیش تعیین شده باشد، امکان ترکیب مهارت‌های مختلف اخلاقی آنها در راستای دست یابی به اخلاقمندترین عامل‌ها فراهم می‌شود(۳۰). استفاده از مدل شناختی شبکه‌های عصبی[۲۶] که با استفاده از ابزار شبکه های عصبی مصنوعی و یک مجموعه آموزشی شامل حالت های مختلف اخلاقی ماشین برای آموزش سیستم، امکان نتیجه گیری در حالت‌های مختلف اخلاقی برای عامل‌ها فراهم می شود؛ و استفاده از روش‌های مختلف ترکیبی که شامل ترکیب روش‌های مختلف دست یابی به عامل‌های اخلاقی است؛ تا استفاده از نظام ارزش ها[۲۷] که مبنی بر قواعد طلایی[۲۸] بوده و شامل مجموعه ای از ارزش های اخلاقی برای ماشین است و هم چنین استفاده از فضیلت اخلاق[۲۹] که با تاکید بر رفتار عامل‌ها و نیز ارزش های اخلاقی موجود در آنها به دنبال تصمیم گیری در مورد مسائل اخلاقی مختلف است (۳۰و۳۴و۳۵).

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

چگونه یک عامل اخلاقی مصنوعی می تواند ارزش های انسانی را نمایش دهد؟ در سال ۲۰۰۶ محققین و صاحب نظران در راستای دست یابی به یک رویکرد عملی برای مواجهه با مسائل اخلاقی ماشین به تشریح دو کلاس از عامل‌ها پرداختند: ماشین های موقعیت-کنش[۳۰] 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 پیش بینی شده است) و ماشین‌های انتخابی[۳۱] که قادر به بهینه سازی پیامدهای ممکن بوده و می توانند فعالیت مربوط به بهینه ترین خروجی را انتخاب کنند. دسته اول از ماشین ها می توانند رفتارهای پیچیده ای از خود نشان دهند، اما از آنجا که تنها اهداف ضمنی[۳۲] دارند، رفتارهای غیر قابل انعطافی داشته و به سختی در محیط ها 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 می کنند (۳۶).

محققان اخلاق ماشین به این که هر عامل مصنوعی اخلاقی می تواند یک عامل اخلاقی ضمنی باشد اتفاق نظر دارند، که می تواند همه اهداف مورد انتظار خود را تحت یک رفتار کاملا 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 ویژگی های کلیدی معینی در عامل‌های اخلاقی، ثابت در نظر گرفته شوند و توسعه در این زمینه به سمت تشکیل اهداف ضمنی برای ماشین‌ها و دست یابی به ماشین‌های موقعیت-کنش گام بر دارد. بایستی اذعان داشت که اتخاذ این فرایض، شباهت کمی به کاربرد عامل های ابرهوش داشته و حتی معیار غیر محلی بودن برنامه ریزی سیستم نیز ارضا نمی شود. هر یک از این فرایض به نوعی با تعاریف و بایدهای مفهوم ابرهوش در تعارض بوده و چالش هایی را در پی دارند. البته بایستی توجه داشت که با این حال حذف هر یک از آنها نیز چالش‌های طراحی جدیدی را می طلبد. برخی از این فرایض و چالش های پیش روی آنها عبارتند از:

فرض: عامل های مصنوعی اخلاقی توسط انسان ساخته شده اند و نمی توانند 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 (۲۵و ۳۷)، که این امر اساساً در تعارض با تعریف ارائه شده برای ابرهوش است.

فرض: انسان بسیار قدرتمندتر از عامل های مصنوعی اخلاقی است لذا استفاده از تئوری‌های بازی برای اجبار عامل به همکاری امکان پذیر است.

چالش طراحی: این فرض نیز با نادیده گرفتن توانایی های 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: عامل های مصنوعی اخلاقی می توانند در محیط هایی که بسیار به محیط طراحی خود شبیه هستند مورد تست و آزمایش قرار گیرند (۳۰).

چالش طراحی: این فرض نیز با تحدید کاربردهای عامل های هوشمند و نادیده گرفتن توانایی‌های آن ها، در راستای ماشین های 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 سیستم هایی که پیش تر تولید شده اند ممکن است دارای رفتارهایی باشند که برای بشر امروزی پر مخاطره باشد (۲۵). به تعبیر دیگر در صورتی که عامل‌های مصنوعی اخلاقی تنها برای محیط هایی با محرک‌ها و شرایط خاص طراحی شده باشند، تضمینی از رفتار امن آنها با قرارگیری در محیط هایی 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 می کند (۳۰).

چالش طراحی: با توجه به توانایی‌های منطقی بالای عامل‌های هوشمند همان طور که پیش تر ذکر شد، دست یابی به ابرهوش

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

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 توانایی های چشم گیر خود، عامل هایی به مراتب توانمندتر از خود ایجاد کند. به تعبیر دیگر پتانسیل موجود در انفجار هوش می تواند 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 هم چنین توانایی‌های آن، امکان نظارت انسانی در طراحی این عامل‌ها میسر نخواهد بود.

همان طور که مشاهده می شود هر یک از این فرایض که به نحوی در جهت اهداف ماشین‌های موقعیت-کنش اتخاذ شده اند، 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 (۱۹و ۳۸).

نتیجه گیری

با پیشرفت دانش هوش مصنوعی و موفقیت آن در شبیه سازی برخی رفتارهای الگوی انسانی، حتی با حذف برخی از محدودیت های لاجرم موجود در آن نظیر خستگی و امکان اشتباه، شاهد ظهور ماشین هایی هستیم که حتی بهتر از الگوی انسانی قادر به انجام وظایف خود هستند. علیرغم این موفقیت ها بایستی اذعان داشت که هنوز دست یابی به هدف غایی هوش مصنوعی (شبیه سازی رفتار جامع انسانی) میسر نشده است. یکی از مهم ترین موانع دست یابی 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 تک منظوره بوده و به طور کلی می توان شاهد برتری انسان نسبت به ماشین بود. با رفع موانع موجود و تحقق آرمان های هوش مصنوعی، شاهد نسل جدیدی از ماشین ها تحت عنوان ابرهوش خواهیم بود که قادر به طراحی ماشین هایی هوشمندتر از خود هستند. این ماشین‌ها نیز می توانند نمونه هایی هوشمندتر از خود ساخته و همین طور تا آخر در یک چرخه بازخوردی مثبت شاهد طراحی ماشین هایی هوشمندتر از مرحله قبل باشیم. از این واقعه تحت عنوان انفجار هوش یاد می شود.

با افزایش روز افزون وظایف تصمیم گیری ماشین و نیز برتری عقلانی نسل جدید ماشین‌های تولید شده، به منظور تضمین رفتار امن ماشین توجه با مسائل اخلاقی این ماشین ها امری ضروری خواهد بود. در این راستا دو رویکرد برای طراحی ماشین‌های هوشمند پیشنهاد شده است که عبارتند از ماشین های موقعیت-کنش و ماشین های انتخابی. در نوع اول از ماشین ها با پیش بینی همه حالت‌های ممکن به تبیین کنش های مناسب ماشین پرداخته شده است در حالی که در نوع دوم با در دسترس بودن مجموعه ای از کنش ها، ماشین قادر به انتخاب مناسب ترین کنش و بهینه سازی خروجی است. علیرغم این که ماشین های موقعیت-کنش در تعارض با مفهوم ابرهوش هستند، اما برای تضمین رفتار امن ماشین بایستی در این راستا گام برداشته شود.

واژه نامه

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

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

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 هایی 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

1. Moravec H (1998). When will computer hardware match the human brain. Journal of Evolution and Technology. 1: 1-12.
2. The Blue Brain Project website. Available at: http://bluebrain.epfl.ch.htm. Accessed: 25 Nov. 2012.
3. The Institute of the Ecole Polytechnique in Lausanne website. Available at: http://www.epfl.ch/index.fr.html. Accessed: 25 Nov. 2012.
4. Russel S, Norvig P (2010). Artificial Intelligence: A modern approach. 3rd edition. Pearson Education. New Jersey. pp: 26-58.
5. Stahl BC (2004). Information, Ethics, and Computers: The problem of autonomous moral agents. Minds and Machines 14: 67–83.
6. Bechtel W (1985). Attributing responsibility to computer systems. Metaphilosophy, 16: 296–306.
7. Stewart I (1996). The interrogator's fallacy. Scientific American. 275: 172–175.
8. 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.
9. Allen C, Varner G, Zinser J (2000). Prolegomena to any future artificial moral agent Journal of 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. 14th 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 پژوهش ها ي علوم انسانی» فصلنامه اخلاق در علوم و فناوری، ۱۳۹۰، شماره ۲: ۶۶–۷۵.

14. Anderson SA (2011).Machine Metaethics. 1st Edition. Cambridge University Press. Cambridge. pp: 21-27.
15. Moor JH (2011). The Nature, Importance, and Difficulty of Machine Ethics. Machine Metaethics. 1st 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. 1st 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 singugical singularity. Presented at the VISION-21 symposium.
23. Bostrom N (2003). Ethical Issues In Advanced Artificial Intelligence. 2nd 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. 1st Edition. Oxford University Press. New York. pp: 15-26.
25. Yudkowsky E (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk.

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]

In Global Catastrophic Risks, Bostrom and Cirkovic, Eds. Oxford University Press. Oxford. pp: 308-345.

26. Kurzweil R (2005). The Singularity Is Near.1st Edition. Penguin Books. London. pp: 1-12.
27. Hall JS (2007). Beyond AI: Creating the Conscience of the Machine. 1st Edition. Prometheus Books. US. pp: 45-59.
28. Posner R (2004). Catastrophe: Risk and Response. 1st 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.1st 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. 1st 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.

[Downloaded from ethicsjournal.ir on 2026-09-16]

[DOR: 20.1001.1.22517634.1392.8.4.4.7]